INVESTIGATING THE RELATIONSHIP BETWEEN STUDENTS' SELF-ASSESSMENT AND RATINGS OBTAINED FROM A FORMAL ORAL PROFICIENCY INTERVIEW (OPI)

A Dissertation

Submitted to the
Faculty of Argosy University Campus
College of Education

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Education

By

Salem Abdelhamid Elfiky

April 2012

INVESTIGATING THE RELATIONSHIP BETWEEN STUDENTS' SELF-ASSESSMENT AND RATINGS OBTAINED FROM A FORMAL ORAL PROFICIENCY INTERVIEW (OPI)

INVESTIGATING THE RELATIONSHIP BETWEEN STUDENTS' SELF-ASSESSMENT AND RATINGS OBTAINED FROM A FORMAL ORAL PROFICIENCY INTERVIEW (OPI)

A Dissertation

Submitted to the
Faculty of Argosy University Campus
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Education

By

Salem Abdelhamid Elfiky

Argosy University

April 2012

Dissertation Committee Approval:

_____ _____
Dissertation Chair: Dr. Scott Griffith Date

_____
Committee Member: Dr. Gordon Jackson

_____ _____
Committee Member: Dr. Barbara Cole Program Chair: Dr. Ardella Dailey

INVESTIGATING THE RELATIONSHIP BETWEEN STUDENTS' SELF-ASSESSMENT AND RATINGS OBTAINED FROM A FORMAL ORAL PROFICIENCY INTERVIEW (OPI)

Abstract of Dissertation

Submitted to the
Faculty of Argosy University Campus
College of Education

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Education

By

Salem Abdelhamid Elfiky

Argosy University

April 2012

_____
Dissertation Chair: Dr. Scott Griffith

_____
Committee Member: Dr. Gordon Jackson

_____
Committee Member: Dr. Barbara Cole

Department: College of Education

ABSTRACT

INVESTIGATING THE RELATIONSHIP BETWEEN STUDENTS' SELF-ASSESSMENT AND RATINGS OBTAINED FROM A FORMAL ORAL PROFICIENCY INTERVIEW (OPI)

The purpose of the study was to develop and validate a self-assessment instrument that can be used to obtain a highly reliable estimate of the foreign language speaking proficiency of native speakers of English. A descriptive research design was employed using a survey to collect data from students who had completed 63-week basic courses in Category-IV languages (Arabic, Chinese and Korean) at the Defense Language Institute Foreign language Center (DLIFLC) in Monterey, California.

The primary data collection method was self-assessment via questionnaire. The questionnaire employed in this study was a Can-Do-Scale (CDS) that was designed to enable native speakers of English to assess their foreign language in speaking. Participants in the study were asked to respond to 30 CDS items related to how well they could perform speaking tasks in real-life situations. The instrument was designed to measure foreign language speaking ability from Level 0 to 3, including plus levels, on the Interagency Language Roundtable Scale.

A Spearman's rho correlation between CDS and OPI ratings was statistically significant ($r = .272$, $p < .05$). Although the correlation was low, in the researcher's opinion, it was strong enough to indicate that the CDS can be used as a tool for diagnostic assessment purposes to identify students' strengths and weaknesses. The relationship between CDS and OPI ratings was also examined by calculating the extent to which the ratings were the same. The percentage of perfect agreement between CDS

and OPI ratings was 58.3%, and in the case of the remaining students, who had discrepant ratings, the majority (34%) of the ratings were only a plus level apart. In other words, 92.3% of the students in the sample had a CDS rating that was either exactly the same as their official, end-of-course OPI rating or very close to it, i.e., only a plus level above or below the OPI rating.

Stepwise and simultaneous multiple regression analyses showed that the predictors of exact agreement between CDS and OPI scores were having an OPI score of level 2 or higher (p. = .000), gender (p. = .014) and military branch (p. = .047), indicating that this constellation of variables had the highest association in this model with the level of agreement between CDS and OPI.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I am truly and deeply indebted to everyone who supported me throughout my dissertation process. It would not have been possible to write this doctoral thesis without the help of the kind of people around me. Above all, I would like to thank my wife and children for their constant personal support and great patience. My brothers and sisters have given me unequivocal support throughout, for which my mere expression of thanks likewise does not suffice.

This thesis would not have been possible without the help and support of my committee supervisor, Dr. Scott Griffith. His leadership, support, attention to detail, hard work, and scholarship have set an example I hope to match someday. The good advice, support and friendship of my second committee member, Dr. Barbara Cole, has been invaluable on both an academic and personal level. My special thanks to my colleague and committee member, Dr. Gordon Jackson, who never got tired of reviewing my work; his inestimable feedback enabled me to gain a better understanding of the dissertation writing process.

I am most grateful to Dr. Donald Fischer, who supported the study academically, logistically, and financially. I want to express my sincere appreciation to Dr. Gary Hughes, who devoted an enormous amount of time to help me with the statistical analysis. I want to thank Dr. Thomas Parry, Mr. Deniz Bilgin, Mr. James Dirgin, Dr. John Lett, Dr. Martha Herzog, Dr. Mika Hoffman and Dr. Jeffrey Crowson, who helped me with the survey validation process.

I want to thank all my colleagues in the Proficiency Standards Division (PSD) for their diligent effort, support and encouragement. They contributed toward many aspects

of this dissertation, such as quality control and third rating of the OPI tests, even on their own time.

For any errors or inadequacies that may remain in this work, the responsibility is of course entirely my own.

DEDICATION

I would like to dedicate this dissertation to my first teacher, my father, Abdelhamid Elfiky, and my stepmother, Wedad Mohamed who passed away in Egypt without my having the opportunity to say goodbye to them. My father taught me at a young age morality, respect and the love of God. He taught me the value of an education and inspired in me the drive and determination to pursue my doctoral degree. He was more than a father after I lost my mother at the age of two; he became a father, a brother, a friend, and everything in my life.

My stepmother raised me from the age of two and became more than my biological mother to me. I know how important it was for you to see me graduate, but unfortunately you were unable to do so. You will be in my heart all my life. I thank you for your sacrifices and everything you did for me, but especially for your unconditional love and friendship.

CHAPTER ONE

Problem Statement

The United States is suffering from a deficiency in foreign language competence in comparison with other western countries (Marmolejo, 2010). The lack of language capabilities among United States troops created significant challenges to their ability to connect with the public during the war in Iraq, and the same problem has happened in Afghanistan. Kruse and McKenna (2008) reported "The military's lack of language skills and cultural expertise is a symptom of the larger problem facing the nation as a whole."

In October 2001, the United States launched the Global War on Terror (GWOT), an international military operation led by the United States and the international community against the terrorist group al-Qaeda. It began in Afghanistan in response to the attack on the New York World Trade Center and the Pentagon in Washington by the al-Qaeda organization on September 11, 2001. As a result of this attack, approximately 2800 people lost their lives in New York, and hundreds more died in Washington and Pennsylvania. The exact number may never be known (9/11 commission report, 2002). GWOT continued in March 2003 when the United States invaded Iraq and deposed Iraqi President Saddam Hussein and his regime (Belasco, 2011). The Congressional Research Service (CRS) released official, unclassified statistics from the Department of Defense (DoD) reporting that United States military personnel in Iraq totaled 141,300 as of March 2009 (Schwartz, 2009).

McFate (2004) stated “Although “know thy enemy” is one of the first principles of warfare, our military operations and national security decision-making have consistently suffered due to lack of foreign cultures.” Due to the lack of cultural awareness, military personal are at risk of losing their lives (Wildes, 2005). A recent report by CRS (2010) provided statistics on military deaths, divided between the period of major combat operations (March through April, 2003) and the ongoing presence of U.S. forces in Iraq after the end of major combat operations (May 1, 2003 through February 6, 2010). The total number of deaths in Iraq was reported to be 4,365, while the total number of individuals wounded was 47, 224. Some individuals died or were wounded in non-hostile combat due to the lack of foreign language skills and cultural awareness (Fischer, 2010). “It is important to respect Iraqi culture and religious practices. Iraqis who feel that the foreign troops respect them might provide information about the insurgency, and that could save American lives” (Wildes, 2005).

There are thousands of soldiers around the world who have some language proficiency, and they need to be assessed to ensure that they can carry out speaking tasks assigned according to their levels of speaking proficiency. Glenn Nordin stated, “Develop a cadre of military linguists with high-level linguistic skills that will be a “core competency” in fighting the Global War on Terrorism”

The Defense Language Institute Foreign Language Center (DLIFLC) is the main institution for foreign language testing within the Department of Defense. DLIFLC testers work on collateral duty because their main job is teaching. Testers are sufficient to test DLIFLC students and employees of federal agencies, but not to carry out a large scope of testing load for U.S. military personnel all over the world, especially post 9/11

for people in combat zones. According to Barcinas (2011), DLIFLC does not have enough testers to fulfill the increasing demands of evaluating speaking proficiency for thousands of linguists.

In combat zones, commanders need to have up-to-date knowledge of the speaking proficiency of all their linguists to be able to assign them to speaking tasks that are commensurate with their ability, which may save lives (Barcinas, 2011). In addition to the number of people to be tested, time zone differences would be a challenge to using face-to face or telephonic Oral Proficiency Interviews (OPI) to accomplish this assessment task because testers are in one location and people to be tested are in a variety of time zones.

Background of the Study

According to a Lockheed Martin report (2010), cultural competence plays a significant role in dealing successfully with a diverse political, multicultural and multinational environment. “Succeeding against today’s enemies - who may not wear uniforms, who are highly adaptive and who are able to slip into environments and cultures foreign to our operators - requires specialized skills and training beyond conventional military capabilities” (Lockheed Martin, 2010). “The complexity of today’s and tomorrows’ strategic environments requires that Special Operation Forces (SOF) operators maintain not only the highest levels of war fighting expertise but also the cultural knowledge and diplomacy skills” (Olsen, 2009).

The Defense Language Institute Foreign Language Center (DLIFLC) at the Presidio of Monterey, California, became the centerpiece of the Department of Defense’s (DOD) post-9/11 language transformation effort. This effort identifies the need to

increase foreign language skills and cultural knowledge in breadth and depth across all DOD departments, with the ultimate goal of all employees having at least some foundation-level, second language capability.

DLIFLC has provided foreign language education, evaluation, and sustainment for all military services, Army, Air Force, Navy, and Marine Corps, for more than six decades. DLIFLC is the largest foreign language teaching institution in the world and provides language instruction in 26 languages and recruits native speaker instructors from all over the world (DLI catalog, 2011). DLIFLC's goal is to ensure that students meet the requirements of the respective military services that have assigned them to study a foreign language. The language programs must meet high standards so that functional language skills may be developed for professional use in real-world communication situations. The teaching methods are task-based, learner-centered and proficiency-oriented, employing authentic materials. The instructional programs at DLIFLC are responsive to the foreign language needs of a wide variety of military jobs in the world.

Admission to DLIFLC is limited to members of the armed forces either on active duty or in a reserve component and civilian employee of the DoD or other federal agencies. Each student must be sponsored by his/her service or employing agency. The agency directs which foreign language the individual will study. Usually, before a student is selected for a language program, a specific vacancy requiring foreign language skills must exist. Each candidate for the basic program must be a high school graduate and have taken the Defense Language Aptitude Test (DLAB), which measures aptitude to learn a foreign language.

The language programs at DLIFLC are divided into four categories based on the difficulty of the language for native speakers of English. Students who score 95 on the DLAB will study a language from Category I, such as French or Italian, for 26 weeks. Students who score 100 will study a language in Category II, such as German or Indonesian for 36 weeks (DLI catalog, 2011). Students who score 105 will study a language in Category III, such as Farsi or Turkish for 47 weeks. Students who score 110 will study a language in Category IV, which includes the most difficult languages for native speakers of English. Arabic, Chinese, Korean, Japanese and Pashto are Category IV languages that have the longest language programs at DLIFLC: 63 weeks.

DLIFLC uses the Interagency Language Roundtable Skill Level Descriptions (ILR) as a curriculum guide for academic activities. The ILR describes six base levels of proficiency from 0 (No Proficiency) to 5 (Functionally Native Proficiency). These are not exhaustive descriptions of what a person with a given level of proficiency can do. They are holistic statements of overall expected ability.

Each base level represents a range of proficiency and has a “threshold”, if an examinee proves during the Oral Proficiency Interview (OPI) that he or she can perform all the language tasks for a given level and meets the minimum performance criteria required for that level, he or she is said to have “crossed the threshold” for that level. While one examinee maybe just across the threshold of a given base level, another examinee may be solidly within the range of the same level.

As noted above, each level represents a range of proficiency, rather than a point on a scale, and these ILR level ranges increase in size progressively. The scope of tasks and functions that an individual can perform adequately at level 1 is much smaller than

that at level 2 and higher. The ILR describes the minimum performance criteria for each base level range. One of the implications of this is that two examinees may receive the same rating in a language but clearly exhibit different profiles within the same range. Each base level range contains a “plus” level.

The ILR Skill Level Descriptions describe six “base levels” (0 to 5) and five “plus levels” (0+ to 4+). A plus is assigned when an examinee’s proficiency *substantially exceeds* the proficiency associated with one base skill level but does not fully meet the criteria for the next base level. "Substantially exceeds" is operationalized in two basic ways. For example, an examinee who can perform all level 1 tasks satisfactorily and can perform most, but not all, level 2 tasks satisfactorily, substantially exceeds the requirements for level 1, but does not fully meet the criteria for level 2, and is therefore rated a 1+.

A plus may also be assigned when an examinee's performance is very close to the next higher level, but does not cross the threshold of that level. For example, if a level 1 examinee performs all level 1 tasks in such a superior way that he or she is extremely close to level 2, but never actually crosses the level two threshold (as indicated by failed probes to level two), the examinee may be rated 1+.

DLIFLC has multiple proficiency and performance goals associated with each of its instructional programs, basic, intermediate, and advanced. All of these goals build on the minimum proficiency outcomes of the basic language program. Approximately 40 percent of the basic program’s graduates exceed the minimum expectations. The minimum graduation requirement for the basic instructional program in every language

are a U.S. government ILR criteria of level 2 in reading comprehension, level 2 in listening comprehension, and level 1+ in speaking ability. In the intermediate program the minimum graduation requirements are level 2+ in reading comprehension, level 2+ in listening comprehension, and level 2 in speaking ability. In the advanced program the minimum graduation requirements are level 3 in reading comprehension, level 3 in listening comprehension, and level 2+ in speaking ability (DLI catalog, 2011).

As a part of its language transformation program to meet the challenges facing national security after 9/11, DLIFLC implemented the Proficiency Enhancement Program (PEP) in 2005. The main goal of PEP is to help students increase their language proficiency to level 2+ in reading and listening, and to level 2 in speaking ability. The methods of achieving the main goal of PEP are to reduce class size and enhance the use of technology in the classroom. In Category IV languages, including Arabic, Chinese and Korean, class size was reduced from ten students to six. PEP aims to increase target-language practice and the amount of teacher attention received in smaller classes. The effects of using technology such as SMART Boards, Tablet PCs and iPods on foreign language learning and teaching are also being evaluated.

DLIFLC uses the Defense Language Proficiency Test (DLPT) to measure student's proficiency in reading and listening. The fifth generation DLPT5 has been developed entirely based on authentic texts and audio passages. This test creates challenges for students to meet the minimum graduation requirement, particularly at higher proficiency levels, since it requires understanding of the target language culture.

This study focuses on speaking skills only. However, the receptive skills of listening and reading make a significant contribution to the productive skill of speaking.

Krashen and Terrell (1983) believe that the first principle of "the natural approach" to language teaching and language acquisition theory is that comprehension precedes production, i.e., listening or reading comprehension precedes speaking or writing ability (p. 21). Au (1993) indicates that students develop literacy based on "real" reading and listening experiences that allow them to use the language in real-life situations.

At the DLIFLC speaking proficiency is measured by the Oral Proficiency Interview (OPI) which is a task-based test with conversational aspects in which the examinee speaks with two trained testers for 20 to 45 minutes. The test is a simulation of real-life conversations. At the same time, the test is designed to gather sufficient data about the examinee's speaking abilities in the target language to permit the accurate matching of that examinee's performance to the ILR rating criteria for the OPI. The test is designed to be highly flexible, making it possible for the testers to find the limits of the examinee's proficiency by using different kinds of tasks, topics, and "role play" situations during the test (OPI 2000, 2010). Every action by a member of the testing team should serve one of two purposes: (1) to identify the kinds of tasks that the examinee is able to carry out successfully in the language (establishing the "floor" of the examinee's proficiency), and (2) to identify tasks that cause the examinee's performance to break down and prevent him or her from sustaining the next higher level (establishing the "ceiling" of the examinee's speaking proficiency). To put it another way, the purpose of testing is to identify what the examinee "can do" and "cannot do" in the language.

While the basic structure of different OPIs is the same, the nature and complexity of the tasks and topics will vary greatly from one test to another and also between levels. The OPI measures "general proficiency" which is the ability to accomplish real-world

communication tasks, including social conversation and work-related tasks relevant to the individual, either within the target culture or during encounters with individuals who speak the target language. Examinees with higher proficiency levels can accomplish tasks of increasing complexity. Proficiency is unrelated to how or where the examinee acquired the language (OPI 2000, 2010).

Purpose of the Study

The purpose of the study was to develop and validate a self-assessment instrument that can be used to obtain a reliable estimate of the foreign language proficiency of native speakers of English. To realize this purpose, the researcher carried out the following steps: 1) investigate the relationship between two types of measures of oral proficiency: level scores inferred from the self-assessment instrument and ratings obtained from a formal Oral Proficiency Interview (OPI). 2) Investigate the impact of various variables that the literature suggests are likely to affect the validity, reliability and accuracy of self-assessment scores and how well students assess their speaking proficiency in a second language or a foreign language.

Theoretical Foundations

This study examines the relationship between two types of measures of oral proficiency. Several theories provided a framework for this study because they are directly related to self-assessment. The theories that provided a framework for this study were the constructivist theory, the theory of multiple intelligences, and the social-cognitive theory (DeMent, 2008).

Constructivism as a theory of learning, or psychological constructivism, emerged from the work of cognitive psychologists such as Piaget, Vygotsky and Bruner. Bruner (1966) views learning as an active process in which learners construct new ideas or concepts based upon their present or previous knowledge. Bruner (1966) describes learning as a procedure in which learning is constructed by students rather than given to them. This happens through an individual's own interaction with a set task. Bruner (1966) states that a theory of instruction should address four major aspects:

(1) Predisposition towards learning.

(2) The ways in which a body of knowledge can be structured so that it can be most readily grasped by the learner.

(3) the most effective sequences in which to present material.

(4) The nature and pacing of rewards and punishments.

Education is most powerful when learners discover new information and apply it in real-life situations (Bruner, 1960). Students play an active role in examining, conducting and monitoring their own learning, but teachers' serve as facilitators (Murphy, 2005).

Howard Gardner wrote about the multiple-intelligences theory in 1983. Developing student's ability to understand themselves better is the core of self-assessment and that is the reason that multiple-intelligences theory has been selected in this study (p. 116). Gardner defined intelligence as "the ability to solve problems or to create products that are valued within one or more cultural settings" (1999, p. 33). Gardner focuses on the aspect of intrapersonal learning, which is the main part of self-evaluation. Regarding the existence of multiple intelligences, Gardner states: "Individuals

have a number of domains of potential intellectual competencies which they are in the position to develop if they are normal and if the appropriate stimulating factors are available" (1983, p. 284). The multiple intelligence idea of assessment is authentic measures which allow students to show what they have learned in a setting that matches the environment in which they expect to show that learning in real life (Armstrong, 1994). Gardner formulated a list of seven intelligences associated with scientific and mathematical thinking:

(1) Linguistic Intelligence involves having a mastery of language.

(2) Spatial Intelligence gives one the ability to manipulate and create mental images in order to solve problems.

(3) Musical Intelligence encompasses the capability to recognize and compose musical pitches, tones, and rhythms.

(4) Bodily-Kinesthetic Intelligence is the ability to use one's mental abilities to coordinate one's own bodily movements.

(5) Interpersonal Intelligence is the ability to notice and make distinctions in the moods, intentions, motivations, and feelings of other individuals.

(6) Intrapersonal Intelligence is the ability to act adaptively on the basis of self-knowledge. This intelligence includes having an accurate picture of one’s strengths and weaknesses, the ability of self-understanding (Armstrong, 1994).

(7) Environmental Intelligence is the ability to discern similarities and differences and make classifications among the living organisms in one’s environment (Gardner & Hatch, 1989).

The third theory in this study is the social-cognitive theory, which states that human achievement depends on the interactions among one's behaviors, personal factors and environmental conditions (Bandura, 1986). This theory indicates that if an individual would like to learn to do something, he or she may have a good opportunity to succeed in doing so by observing, and then imitating the particular action involved. During the imitation process, the person is learning while he or she attempts to complete the activity on his or her own.

The cognitive processes play a central role in education. Consequences and events drive behavior (Bandura, 1969). Bandura believes that people learn through observing others' attitudes, behaviors, and the outcomes of those behaviors. He states that four processes lead to learning:

(1) Attention Processes.

(2) Retention Processes.

(3) Motor Production Processes.

(4) Motivational Processes.

According to Bandura's (1986) social-cognitive theory, students' beliefs about their academic capabilities, or self-efficacy beliefs, are good predictors of their academic achievement and of their next career choices and decisions.

Hypotheses / Assumptions

H1: There is a relationship between student scores on a self-assessment survey and those on a criterion-referenced speaking proficiency test.

$H1_0$: There is no relationship between student scores on a self-assessment survey and those on a criterion-referenced speaking proficiency test.

H2: There is an impact of various variables on the psychometric properties (validity, reliability and accuracy) of student self-assessments of second language speaking proficiency.

$H2_0$**:** There is no impact of various variables on the psychometric properties (validity, reliability and accuracy) of student self-assessments of second language speaking proficiency.

Research Questions

RQ1: What is the relationship between student scores on a self-assessment survey and those on a criterion-referenced speaking proficiency test?

RQ2: What is the impact of various variables on the psychometric properties (validity, reliability and accuracy) of student self-assessments of second language speaking proficiency?

Definition of Terms

1. Achievement Test**:** A test designed to measure the knowledge of an individual in something that has been learned or taught, e.g., arithmetic or typing.
2. Concurrent validity: Is demonstrated where a test correlates well with a measure that has previously been validated. The two measures may be for the same construct, or for different, but presumably related, constructs
3. Criterion–Referenced Test: A test in which the results can be used to determine a student's progress toward mastery of a content area. Each student's performance is compared to an expected level of mastery in a content area rather than to other students' scores.

4. DLPT5: The fifth generation of the Defense Language Proficiency Test.
5. Face Validity: The appearance a test gives of measuring what it is designed to. The OPI has a high degree of face validity because it tests speaking by having people speak.
6. ILR Speaking Skill Level Descriptions: This document (see Appendix C) is a scale that is designed to measure language proficiency and describes different levels or degrees of speaking proficiency, ranging from 0 (No Proficiency) to 5 (Functionally Native Proficiency**).** These are not exhaustive descriptions of the communicative tasks that a person with a given level of proficiency can perform. They are holistic statements of overall ability.
7. Inter-Rater Reliability: The extent to which two or more raters obtain the same result when using the same instrument to measure a concept.
8. Language Proficiency Test: A test designed to measure a learner's ability to function in a language in real life regardless of the type of education he/she may have had in the language.
9. Level Checks: Start after the warm-up at the examinee's assumed working level (floor). These tasks establish what an examinee *can* do in the language.
10. Master Testers: A group of highly trained and experienced oral proficiency testers.
11. Military linguist: A military person who is skilled in at least one foreign language in addition to his/her native language.
12. Norm-Referenced Test: A test referenced to norms based on the performance of students across the nation which is designed to compare student achievement relative to other students' achievement.

13. OCONUS Immersion: Some of DLIFLC's best performing students may have an opportunity to travel on a three to six week long in-country (OCONUS) immersion.
14. Oral Proficiency Interview (OPI): A standardized, criterion-referenced procedure for the global assessment of functional speaking ability. The interviewee's performance on speaking tasks is compared with the criteria for each of ten proficiency levels described in the ILR.
15. Oral Proficiency: The capability to carry out real-world communication tasks in speaking.
16. Reliability: Refers to the degree to which a test generates consistent results.
17. Probes: Establish what an examinee *cannot* do (ceiling) in the language.
18. Role-Play: A special task that enables testers to check types of language use not easily tested in a conversational format.
19. Rubric: A guideline for rating student performance.
20. Testers: Teachers who are certified by the DLIFLC's Proficiency Standards Division (PSD) to test speaking for basic-course students.
21. Threshold or Base Level: Any place or point of entering or beginning: the threshold of a new level.
22. Warm-Up: First phase of an OPI. Nothing the examinee says during the warm-up is rated. Instead, it allows the testers to set the stage for the rest of the test.
23. Wind-Down: Final phase of the OPI. Nothing the examinee says during the wind-down is rated, because the language sample has at this point already been collected. Instead, it gives the testers the opportunity to conclude the test on a positive note.

Limitations

1. Participants in the study are military linguists in four military services.
2. The study is limited to military personal ranked from newcomers to senior officers.
3. The self- assessment questionnaire was based upon the Interagency Language Roundtable Skill Level Descriptions (ILR).
4. Not all the students who participated in this study had travelled to the target country where the language is spoken for an immersion program.
5. The study is limited to students in category-IV language programs.
6. The OPI is a qualitative assessment expressed as a numerical score.
7. The inter-rater reliability may not be perfect.
8. The population of the study was consisted of basic-course students, and their proficiency was limited to level 1+ and 2.
9. The self-assessment questionnaire is written in English, and for some students English is not their first language.
10. The majority of the participants in the study are monolinguals with no prior experience studying their assigned foreign language.

Delimitations

1. Population and sample: The results may not be applicable to other institutes or universities.
2. The findings of the study may not be applicable to other bilingual programs.
3. The findings may not be applicable to students in languages in categories I, II or III.

Significance of the Study

This study has the potential of making a valuable contribution to the assessment of speaking proficiency using the ILR scale. Testers and teachers at the DLIFLC and other government agencies that use the ILR scale can be given the CDS and its scoring rubric to use to evaluate their own proficiency and/or that of their students. Rubrics have become popular with teachers as a way of communicating expectations for homework assignments, providing focused feedback on work in progress, and grading final products (Andrade, 2000; Moskal, 2003; Popham, 1997). Students' participation in self-assessment can help them in becoming more thoughtful and aware of their language proficiency (Tompkins, 2004; Leal, 2006).

The educational profession may benefit from the results of the study by having a better understanding of self-assessment and its implementation in the classroom. The CDS can be used for both summative and formative evaluation. It can be administered as an assessment tool at the end of each semester to estimate students' speaking ability at that point. Teachers may also use the CDS periodically to diagnose students' strengths and weaknesses and plan instruction accordingly. Students themselves could use the CDS to evaluate how well they perform a given ILR language task or function. The frequent use of the CDS and ILR descriptions by teachers, and students can enhance understanding of the criterion requirements to reach each level of proficiency and what a student needs to accomplish to achieve a certain level. The CDS can be valuable to students because they can have an evaluation of their speaking skills that is probably close to, if not exactly the same as, their true speaking ability as measured by a formal OPI. Self-assessment can provide teachers with valuable information regarding students'

perception of their ability at present. Teachers can use this information to identify the needs of the students and prepare tailored instruction and homework materials.

DLIFLC has trained approximately twenty five percent of the faculty to be OPI testers. Those teachers who are not certified OPI testers sometimes give inflated speaking ratings to students since they are not familiar with the ILR criteria. However, students trust the teachers as experts in the field and they tend to believe that teachers' evaluations of their speaking are correct. Hence, students unfortunately neglect to practice speaking.

To address this issue, training on the CDS could be conducted for the teachers who are not certified OPI testers; after the training the teachers would be able to explain the skill-level descriptions to their students and tell the students which CDS items are associated with which level. The major benefit of the CDS is showing students the kinds of tasks they are expected to be able to perform at different levels, and I suspect that would inevitably lead to telling them which items are associated with which level. The CDS would motivate students to aim toward self-improvement and shift their focus from grades to learning.

This study could have a great impact on the profession of foreign/second language teaching. The existence among learners of increased productivity and autonomy, higher motivation, less frustration, and higher retention rates would take place by having tangible evidence of progress when using self-assessment (Ellis, 1994; Gardner & McIntyre, 1991; O'Malley & Pierce, 1996; Rivers, 2001). The language programs investigated in this study are intensive, and students receive six to seven hours of classroom foreign language instruction per day in three skills: reading, listening, and speaking, in addition to final learning objectives (FLO) such as translation, interpretation

and transcription. By the end of the Basic Course some students become frustrated, discouraged, and sometimes stressed to the point that their learning suffers. When students are constantly stressed or if they are in situations that are too stressful, their ability to learn is impaired and recall of previous learning is weakened as well.

This study developed and validated a self-assessment instrument for measuring proficiency in speaking for students learning Arabic, Chinese and Korean, and it can be used with other languages as well. Language assessment in classrooms are common practice with claims about the potential for student-involved assessment in general and rubrics in particular to increase students "self-efficacy and, as a result, lead to improvements in learning and achievement" (e.g., Arter & McTighe, 2001; Quinlan, 2006; Stiggins, 2001).

The implementation of a reliable self-assessment tool can assist students in examining their own progress and play an active role in their education. Students can develop critical thinking skills, which is a part of the evaluation process. Students' active participation in education promotes students' learning when they have tangible evidence of their progress, and it is hoped that this can result in a lower drop-out rate among military students in Category-IV language.

In the aftermath of 9/11 DLIFLC plays an increasingly important role in training military personnel in foreign language skills (DLI catalog, 2011). The "Defense Language Transformation Roadmap" notes that language skill and regional expertise are not valued as defense core competencies, yet they are as important as critical weapon systems (McFate & Jackson, 2005).

As a result of this research, there could be enormous changes in evaluating military linguists in the field. Having an alternative tool to assess speaking in foreign languages could enable the military services and U. S. government agencies to plan missions for their staff based on a reasonably close estimate of their ability. Furthermore, leaders and commanders could project the need for linguists based on future plans. The military services assign different jobs to their staff, and each job requires different foreign language skills, therefore the requirements vary from one agency to another. Military linguists can be assigned jobs by their military units that are commensurate with their ability in their foreign language in reading, listening and speaking.

One of DLIFLC's goals is to ensure that linguists sustain and improve the level of language skills that they have when they graduate from the language programs. The Army requires that its linguists take the listening and reading DLPT for their language on an annual basis. Linguists can take the listening and reading tests through a secure military website and commanders can receive scores in less than 24 hours to update proficiency records for their staff. However, the primary challenge in evaluating speaking is how to evaluate the speaking ability of military linguists who are located all over the world. Conducting a face-to-face OPI requires two certified testers and an examinee to be located at the same place. The OPI is conducted via telephone or video teleconferencing if the examinee is located abroad. This study can have enormous implications for evaluating thousands of military linguists in the war zones and around the world by using the CDS to verify whether or not soldiers and staff sustain their foreign language proficiency.

Evaluating the speaking proficiency of tens of thousands of linguists worldwide by scheduling oral proficiency interviews by two certified testers is not feasible because of the time involved in scheduling, conducting, and reporting the scores of, the interviews. Commanders need to know the current speaking proficiency scores of all their linguists at a moment's notice in combat zones. Time zone differences would be another hindrance to accomplishing this task via the OPI. The time-consuming task of evaluating the large number of linguists in the 26 languages that DLIFLC is responsible to test is almost impossible to finish annually. This is problematic for military linguists who have highly classified jobs and are expected to take the OPI test annually or semi-annually to maintain an up-to-date record of their language proficiency.

The development of a quickly and easily administered, indirect, but valid measure of speaking proficiency would enable the DoD to have an up-to-date record of its employees' level of proficiency. Commanders could assign language jobs to individuals with the requisite level of communicative proficiency, which in turn could save lives in a war zone.

CHAPTER TWO

Literature Review

The purpose of the study was to develop and validate a self-assessment instrument that can be used to obtain a reliable estimate of the foreign language proficiency of native speakers of English. To realize this purpose, the researcher carried out the following steps: 1) investigate the relationship between two types of measures of oral proficiency: level scores inferred from the self-assessment instrument and ratings obtained from a formal Oral Proficiency Interview (OPI). 2) Investigate the impact of various variables that the literature suggests are likely to affect the validity, reliability and accuracy of self-assessment scores and how well students assess their speaking proficiency in a second language or a foreign language.

Theoretical Foundations

Social Constructivism Theory

Many learning theories are linked with self-assessment. The theories that presented a foundation for this study included the constructivist theory, the multiple intelligences theory, and the social-cognitive theory. A major idea of Bruner's constructivist theory is that learning is a dynamic process in which learners create new ideas or concepts based upon their current/past knowledge (DeMent, 2008). The learner selects and transforms information, constructs hypotheses, and makes decisions, relying on a cognitive structure to do so. Cognitive structure (i.e., schema, mental models)

provides meaning and organization to experiences and allows the individual to "go beyond the information given" (Bruner, 1990).

The constructivist view assumes the active role of the learner and that it is intrinsic in humans to construct meaning from experience. Self-awareness develops when the child reaches the level of what Piaget calls the cognitive level of “formal operations.” Flavell (1979) describes the level of formal operations as metacognition where the cognitive functions are monitored and controlled. The constructivist view of learning states that the learner dynamically interprets knowledge from the surrounding world and in interaction with others (Oscarson, 2009). Glasersfeld (1995) stated “all knowledge is instrumental and meaningless in isolation” (p. 177). Consequently, Williams and Burden (1997) say, “education becomes concerned with helping people to make their own meanings” (p. 51). There is no such thing as absolute knowledge. Different individuals will have different understandings of experiences and make meanings that are personal to them when knowledge is internal and personal to the individual. Being aware of one’s own learning should then foster both better and autonomous learning (Oscarson, 2009)

In a student-centered classroom, students apply the constructivist theory by organizing information, exploring the learning environment; the students participating in learning activities monitor their own learning (DeMent, 2008). Teachers play a more supportive and reflective role, and the students are free to explore meaning on their own and engage in critical thinking (Iran-Nejad, 1995).

According to Erwin (1991) the definition of self- assessment is “…the systematic basis for making inferences about the learning and development of students. More specifically, self-assessment is the process of defining, selecting, designing, collecting,

analyzing, interpreting, and using information to increase students' learning and development."

The structure of learner-centered approaches has a close relationship to constructivist theory and both would have the potential of helping students to develop the skills needed to evaluate their own performance. This potential can be implicitly seen in Carless' (2007) conceptual definition of "learning-oriented assessment" (p. 57). According to Carless, a learning-oriented assessment is one whose activities are designed to (1) intentionally connect to learning goals, (2) involve students in the assessment process, and (3) help students improve on future learning opportunities through timely, quality feedback. A number of seminal authors make it clear that these learning-oriented characteristics are just a few of the motivations for designing and practicing student self-assessment (e.g., Boud, 1991; Fink, 2003; Gibbs & Simpson, 2004).

In a task-based assessment students have to evaluate how well they perform the task according to the rating criteria and what needs to be done to improve their performance. This fits with the main components of self-evaluation because when students practice ideas on their own, they gain a better understanding. (DeMent 2008). When incorporating the constructivist theory in classroom practices, teachers engage in an active dialogue with students and guide them so that they can facilitate their own learning (Bruner, 1960). When students self-evaluate, they focus on where they are and where they want to be; from there, they can construct what they need to learn in order to achieve the desired tasks. This is part of the foundation of the constructivist learning theory (DeMent, 2008).

According to Oscarson (2009), there are different forms of constructivist theory, particularly social constructivism. Students understand knowledge as something that grows and increases in the encounter between the learner and the teacher in a social environment. Knowledge can therefore not be "taught" per se; the teacher can only facilitate and guide the learner on the road to learning. In the constructivist theory there is a need for the learner to be aware of his or her own learning so that the learner is able to regulate and evaluate the learning process him- or herself. The development of metacognitive skills is of importance to the learner-centered approach (Oscarson, 2009). The social constructivist perspective on learning puts the student at the center of the learning process and the metacognitive functions are accorded an important role in individuals' building of new knowledge (Gipps, 1994). The constructivism learning theory, which focuses on knowledge construction based on learner's previous experience, is a good fit for student-centered approaches because it ensures learning among learners (Harman & Koohang, 2005; Hung, 2001; Hung & Nichani, 2001; Koohang & Harman, 2005).

Multiple Intelligences Theory

This study also drew upon the principles of the theory of multiple intelligences proposed by Gardner (1983). Multiple Intelligences focuses on the principle that intelligence is "the ability to gain access to and understand one's inner feelings, dreams and ideas" (Gardner, 1993, p. 20). Gardner (1999) identified eight intelligences in his theory of multiple intelligences. He believed that these intelligences do not operate independently; rather, they are used at the same time as people develop skills or solve

problems. The theory of multiple intelligences addresses the aspects of intrapersonal learning, a core component of self-evaluation (DeMent, 2008).

Intrapersonal intelligence (Gardner, 1995) involves the skill of understanding oneself and to realize one's fears and motivations. It also transfers to inner states of being, self-reflection, and metacognition, which is the development of thinking about thinking and is a main component in learning how to self-evaluate. All of these components are important for self-evaluation. Having this intelligence allows people to have an efficient working model and to use this knowledge to regulate their learning (DeMent, 2008). Developing intrapersonal intelligence means that students are able to think really inside themselves, a process that can be enhanced by self-paced projects and choices. Strategies used to make use of this intelligence are self-checking materials, goal sheets, and journal writing (Gardner, 1999).

Multiple Intelligences (MI) theory increases the assessment arena significantly to incorporate a broad range of possible contexts within which a student can express aptitude in a specific area. Multiple Intelligences theory supports the idea that a student should be able to show proficiency in a specific skill, subject, content area, or domain in any one of a variety of ways. The theory of MI suggests that any educational objective can be taught in no less than seven different ways; so too does it imply that any subject can be assessed in no less than seven different ways. On the contrary, regular tests regularly require students be seated at a desk, where they complete the test within a specific amount of time, and they speak to no one during the test. According to Armstrong (1994), the tests themselves frequently consist of linguistic assessment items that students must answer by filling in bubbles on computer-coded forms (pp.121-123).

These same strategies are at the base of self-evaluation. Goleman (1995) posited that the concept of intrapersonal or emotional intelligence is as important as cognitive intelligence. Emotional/Intrapersonal intelligence establishes how well people use the skills they have, including their intelligence. Individuals who are emotionally skillful understand and manage their own feelings well (Goleman). Developing a sense of self and exploring the affective element in education may be particularly advantageous to talented students (Johnson, 2000) because they have a tendency to have an elevated sense of self and sensitivity.

Additionally, many talented students are perfectionists who can be especially critical of their own abilities, which leads to low self-esteem (DeMent, 2008). This may lead to talented students underachieving academically because they have a sense of failure. In addition, some talented students conceal their true talent in order to fit in with their peers as well as their teachers (Diaz, 1998). By developing a sense of self, talented students, especially females, can actually develop appropriate objectives and make decisions based on deeply held values (Badolato, 1998).

Research has shown a relationship between social and emotional development and the school factors of social status, perception of teacher and peers, class participation, achievement, and self-direction in learning (Katz, 1994). Success or lack thereof, in these areas can be linked to either positive or negative self-concept (Katz). Self-evaluation, which employs the theory of Gardner's (1995) intrapersonal intelligence, may be one way to help talented students accomplish their desired objectives.

Social Cognitive Theory

Many other theories of self-regulation focus on the question of how students evaluate learning both individually and socially. Bandura's (1986) social learning theory, most recently called Social Cognitive Theory, subscribes to the principle that individuals have a system of beliefs about themselves that allow them to be in charge of their actions. Social Cognitive Theory has been influential in research on social factors in self-regulation, which focuses on interdependent personal, behavioral and environmental influences (Zimmerman, 2001, p. 19). An individual's behavior is determined by the interplay between these factors. Behavioral outcomes form future expectations. Self-regulation can be seen as a cyclic process which includes three major phases; consideration, performance or volitional control and self-reflection (Zimmerman, 1998). Consideration includes goal setting, strategic planning, and natural interest. Performance includes attention focusing, self-instruction and self-monitoring. The self-reflection processes are self-assessment, attributions, self-reactions and flexibility and it is thus the practice of self-reflection that is the most influential mediator in human agency (Oscarson, 2009).

Social-cognitive theory states that human achievement depends on the continuous mutual interactions among behaviors, personal factors, and environmental situations (Bandura, 1986). Viewing and modeling the performances, attitudes, and emotional reactions of others is a major emphasis of the theory (Bandura, 1977). Observational learning is a main element because it necessitates the learner to focus attention on something being modeled, observe its characteristics, and then replicate what was seen

(Bandura, 1986). Modeling is essential in teaching students how to self-evaluate, thus developing the key foundations of the theory.

Self-efficacy is a component of Bandura's (1993) social-cognitive theory. Self-efficacy concentrates on how self-perceptions of one's ability to perform a task influence engagement in and successful completion of the task (Bandura, 1993). Self-efficacy develops from tasks that are highly challenging yet achievable. Self-efficacy grows through mastery experiences, vicarious experiences, and interpretation of physiological and emotional states. When students carry on tasks and spend effort to ensure success, their self-efficacy is enhanced (Schunk, 2003). Bandura's theory supports the notion that students need more than intelligence and skill to perform successfully. They also need a strong sense of self to use what they have well and to regulate their own learning process (Bandura).

Finally, self-efficacy beliefs are task-specific and build through normative criteria rather than through comparison with others. Students' self-efficacy in adolescence is challenged by psychological and physical changes that influence students' sense of personal control, which may result in lower self-confidence. This lack of confidence can affect students' writing ability because writing requires one to have confidence in organizing and managing the various tasks involved in the process (Bandura, 1997). Schunk (2003) found that efficacy does not need to be extremely high for effective learning; instead, it needs to be high enough only to sustain engagement in present and future tasks. Schunk added that prerequisite knowledge and skills, along with high self-efficacy, are good predictors of writing achievement.

Self-evaluation in speaking proficiency and writing development are important factors in self-efficacy (Schunk, 2003). Students' positive self-evaluations enhance their self-efficacy because they can observe that they understood the goal and applied a strategy that promoted their success. Low self-evaluations, on the other hand, do not certainly decrease self-efficacy as long as students can reflect on their work and believe that they can succeed. As students continue to work harder and adopt new strategies through self-evaluation, their self-efficacy increases (Schunk, 2003). In addition, confident students are less apprehensive as writers and have more feelings of self-worth about their writing (Pajares, 2003). According to Bandura (1986), the capability that is most "distinctly human" is that of self-reflection, therefore it is a prominent feature of social cognitive theory. Through self-reflection, people make sense of their experiences, explore their own cognitions and self-beliefs, engage in self-evaluation, and alter their thinking and behavior accordingly (p. 21).

Self-assessment

Introduction to Self-assessment

For the past two decades, self-assessment has been increasingly used in a wide range of educational settings due to the shift from teacher-centered to student-centered teaching approaches (De Saint Leger, 2009). The growing interest in "authenticity and instructiveness" to enhance language proficiency and the current trends in learner-centered approaches (Bachman & Palmer, 1996) have increased the interest in expanding the use of second language self-assessment (Bachman, 2000; Calfee & Hiebert, 1990; Hamayan, 1995). Language testers have been inspired to examine whether students are able to make a significant contribution to their own assessment. In the past twenty years,

self-assessment has become a quite popular issue of research in foreign language education. This focus is due to the rising interest in unusual forms of assessment, in which self-assessment plays a role in decreasing the testing workload of teachers.

The use of self-assessment procedures and a variety of alternatives in assessment can be implemented by teachers in the context in which the learning takes place, allowing students to be evaluated on what they usually do in class, encouraging students to get familiar with the standards and rating criteria, and requiring students to perform or do something while providing information about their strengths and weaknesses (Brown & Hudson, 1998). Self-assessment is one form of alternative assessment that allows students to make judgments on their own learning, as well as reflect upon that learning. Many people, and in particular, young students, feel that they are never understood in the sense of what they can do with the language. Obviously, we need a type of assessment that gives the learner a sense of belonging, success, autonomy, independence, empowerment, and mastery over his or her own destiny, while simultaneously affording the learner a clear understanding of what is being taught (McDonald, 2007).

LeBlanc and Painchaud (1985) indicate that there are two factors that contribute to the effective implementation of self-assessment: (1) creating concrete linguistic situations where the learner can self-assess his/her communicative ability, and (2) creating good criteria that will in turn produce good, productive, self-assessment items. Self-assessment appears to be a natural progression in keeping with information about multiple intelligences, the knowledge era, massive globalization, the transformation of modern society, a climate of unprecedented organizational change, and student migration to the twenty-first century (McDonald, 2007).

Dickinson (1992) reported that there is a considerable scope for self-assessment when tests that are designed for assessing proficiency are used as a learner-teacher feedback device. He stated, “The type of assessment that goes on in the classroom is the primary concern with the learning process, indicating to learners the degree to which they have achieved a standard of performance which is adequate for a particular situation (p.33).”

Definition of Self-assessment

Self-assessment is known by a variety of names such as self- evaluation, self-rating, self-testing, and self-appraisal, etc. (Yang & Xu, 2008). Self-assessment may be defined differently depending on what is being assessed. For example, Self-assessment is defined by McMillan & Hearn (2009) “as the method by which students control and appraise the way they think and behave when learning; students are then able to improve their understanding and skills by recognizing their learning strategies.”

Boud (1986) defines self- assessment as “the involvement of students in identifying standards and/or criteria to apply to their work and making judgments about the extent to which they have met these criteria and standards” (p. 5). This is of course an example of criterion-referenced self-assessment in which learners have to gather data about their own progress or performance and compare it against standards or criteria. This process should not be used to establish students’ grades.

Blanche and Merino (1998) defined self-assessment as information that has been provided by learners about their ability and what they can or cannot do with what they have learned in a course.

The field of self-assessment of language proficiency is associated with knowing how, and under what conditions learners and users of a second language or a foreign language evaluate their own proficiency (Oscarson, 1997). Self-assessment, according to (Oscarson, 1997), comes from the realization that effective learning is best achieved if the student is actively involved in all phases of the learning process.

In recent years, self-assessments have increased in popularity and have been used by a number of higher educational facilities around the globe (Butler & Lee, 2006). Self-assessment nurtures students' skills by allowing them to learn information and judge their performance in meaningful rather than rote methods (McMillan & Hearn, 2009). Although there are various benefits when using self-assessments, there are some reliability and validity concerns connected with educational assessments due to the various differences in students' performances and abilities. For instance, students' proficiency in English and how honest he/she answers the questions may affect the validity and reliability of self-assessment.

The Value of Criterion-Referenced Self-assessment

Andrade (2001) recommends that just giving out and explaining a rubric may develop students' knowledge of the criteria for a task and help students produce work of good quality—or it may not. Simply handing out a rubric does not guarantee much of anything. When students are actively involved in using a self-assessment instrument to evaluate their learning, there is noticeable progress in students' performance of the tasks (Evans, 2001).

DLIFLC and many government agencies use the ILR criteria as a guideline for their entire testing activities; including the development of reading and listening tests and

of the Oral Proficiency Interview. The definition of criterion-referenced self-assessment is a process of formative assessment during which students reflect on and evaluate their learning and the value of their work, judge the degree to which they have met the stated goals or criteria, identify strengths and weaknesses in their work, and revise accordingly (Goodrich, 1996; Gregory et al., 2000; Hanrahan & Isaacs, 2001; Paris & Paris, 2001; Andrade & Boulay, 2003).

Although there is no standard definition of self-assessment in the literature, there are several characteristics of self-assessment common to the various definitions. For one, student self-assessment is criterion-referenced (Andrade& Du, 2007). Frederickson and Collins (1989), Wiggins (1998), and Stiggins (2001) argue that the "criteria for student work must be so transparent that students can learn to evaluate their own work the same way their teacher does."

Another characteristic of self-assessment is an emphasis on promoting learning by providing feedback that guides students' efforts and strategies (Adams, 1998; Lewbel & Hibbard, 2001; Paris & Paris, 2001; Horner & Shwery, 2002). A third characteristic is that it is ongoing: self-assessment frequently engages, monitoring, and regulating one's thinking processes and task performances as they take place (Goodrich, 1996; Andrade & Boulay, 2003).

There are also commonalities in the processes of self-assessment described in the literature. With minor variations, academic self-assessment is scaffolded in the following way (Andrade & Du, 2007):

(1) Teachers share the expectations for the ideal performance with students, often by providing a rubric and/or models or examples of student work (e.g., Stallings & Tascione, 1996).

(2) Students prepare drafts of the assignment and formally and/or informally assess their work according to the rubric and/or the examples (e.g., Gregory et al., 2000; Hart, 1999; Hanrahan & Isaacs, 2001; Lewbel & Hibbard, 2001 Paris & Paris, 2000).

(3) Students use the feedback generated by their self-assessments to guide them in making corrective adjustments to their work (e.g., Adams, 1998).

Self-assessment plays a key role in learners' autonomy which leads to learning and academic success and increases in metacognitive engagement (Paris & Paris, 2001; Rivers, 2001). Criterion-referenced self-assessment is a key element of self- regulation and 'self-judgment' with the potential to scaffold other elements, including goal-setting, planning and self-reaction (Andrade & Du, 2007). Vygotsky defined scaffolding instruction as the "role of teachers and others in supporting the learner's development and providing support structures to get to that next stage or level" (Raymond, 2000, p. 176).

Even though the role of self-assessment in becoming an autonomous, metacognitive, self-regulated learner appears to be logical, it has little practical support. Falchikov and Boud (1989) reviewed studies of self-evaluation, or the correlation between self- and teacher ratings. Stallings and Tascione (1996) employed student self-assessment in high school and college mathematics classes and found, among other things, that most of the students checked their work for accuracy more enthusiastically than students in previous classes who were not exposed to self-assessment practices.

In a survey of undergraduates who had used self- and peer assessment, Hanrahan and Isaacs (2001) reported that students noticed advantages of, as well as difficulties with, self- and peer assessment. Advantages included having a better understanding of grading, increased critical thinking, developing and becoming inspired to do better work in order to impress one's peers. Some learners reported problems with self- and peer assessment when they were "not sure of standards" (p. 59).

Andrade and Du (2007) reported on a study that examined students' reaction to criterion-referenced self-assessment and students' attitudes toward self-assessment after they had experience with it. The study also examined the result of gender differences in students' responses to criterion-referenced self-assessment. The study discovered these findings:

(1) Students reported that their attitudes toward self-assessment became more positive as they gained experience with it.

(2) Self-assessment and teacher expectations were constrained and inextricable.

(3) Students felt they could self-assess effectively and were more likely to self-assess when they knew what the teacher expected.

(4) Students self-assessed by checking, revising their work and reflecting on how well the task was achieved.

(5) Students believed there were multiple benefits of criterion-referenced self-assessment.

(6) The transfer of self-assessment processes was spotty.

(7) A tension between expectations and students' own standards of quality were evidence.

(8) No evidence was found of gender differences (Andrade & Du, 2007).

Validity of Self-assessment

The Limited Validity of Self-assessment

Oscarson (1980) stated, “Most learners of another language have a certain capacity for determining their own language ability—provided that they have at their disposal a measuring standard by which they may express their intuitions” (p.13). Although few people dispute the value of self-assessment as one of the cornerstones of autonomous learning, concerns have been expressed about the validity of self-assessment methods used as a basis for making decisions such as selection for placement in class, grading, and certification (Ekbatani & Pierson, 1998). According to Dickinson (1992), self-assessment may not be an adequate tool when evaluation is undertaken for the purpose of awarding recognition of achievement or withholding recognition from a candidate. There are some concerns about learners’ capacity to view their success.

Students who have had experience in self-assessment might have a higher ability to assess their own language skills for placement purposes as objective assessments of language proficiency (LeBlanc & Painchaud, 1985). Because he is influenced by the significance of self-assessment, Janssen-van Dieten (1989) constructed a self-assessment survey along with a traditional test of Dutch-as a second language for use in adult language courses. However, the relationship between performances on the two types of measures showed no consistency. The survey did not show a major trend to increase self-assessment reliably at higher proficiency levels. Additional reports have found small to no statistically significant relationships between self-assessment and other measures of language ability (Wesche, Morrison, Ready & Pawley, 1990). This weakness may be due

to the role of self-confidence when people assess themselves or to the type of rating scale used in the study.

Dickinson (1987) stated “It is very likely that most learners assessing themselves will be biased in their own favor, which could possibly distort the results" (p. 150). The possibility of misrepresentation suggests that, as with any other measure, a single measure is never perfect. Whatever placement is actually done based on a self-assessment measure either needs to be checked against an alternative measure or, some way needs to be developed so that particular placements can be changed whenever incorrect, for whatever reason (Wolochuk, 2009).

There are two concerns that may affect the validity of self-assessment. (1) Raters vary widely in their strictness, and this lack of consistency can account for as much as 40% of test variance (Cason & Cason, 1984), which is the reason for the increase of using two raters in assessment (2) each level is defined by descriptors so that if the raters are normed to assess the same way to reduce the difference in severity between them, the validity of the assessment is still questionable. The reliability of the self-assessment will be hard to achieve as those descriptors that are poorly defined or incorrectly placed on the scale will undermine the rater’s efforts (North, 1993). The concern is well known in the scaling literature and led Thurstone, (1982) to propose that:

> The scale values of the statements should not be affected by the opinions of the people who helped to construct the scale. This may turn out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale. At any rate, to the extent that the present method of scale construction is affected by the opinion of the readers who help sort out the original statement into a scale, to that extent the validity of the scale may be challenged (p.15).

Self-assessment might be quite successful when used for research but not when students are asked to assess their placement into levels of study in a language program (Wolochuk, 2009). Students who would like to be exempt from study might rate themselves significantly higher in a placement situation than they would in a research setting (Brown & Hudson, 1998). The concern is: can students be trusted to be honest in their self-evaluation. It is hard to distinguish between the natural tendency of a learner to interpret results and the deliberate fabrication of results. The discrepancies occur when self-assessment instruments are used in combination with other standardized measures for placement in a language program (Janssen-van Dieten, 1989; Pierce, Swain, & Hart, 1993). Investigating why such discrepancies occur could produce meaningful insights for second-language acquisition (SLA) college educators who may consider using self-assessment instruments in their language programs.

Previous studies indicated other factors that may affect the validity of self-assessment questionnaires, including the wording of the questionnaire, the language skills being assessed, the level of proficiency of the students, and the cultural background of the students (Strong-Krause, 1997). Bachman and Palmer (1989) reported that students found it easier to assess themselves on how difficult they found a task (e.g., how difficult is it for you to read in English?) rather than on their ability to complete a task (e.g., how well do you read in English?). Preparing a self- assessment survey for oral language requires careful wording so that the assessment itself does not become an exercise in reading comprehension (O'Malley & Pierce, 1996).

We should take into consideration students' level of reading proficiency in the language in which the questionnaire is written. Each statement should be expressed in the first person (e.g., "I can order a simple meal") that the students may relate to it directly.

After conducting a meta–analysis of ten studies of second language proficiency Ross, (1998) pointed out some concerns about "can-do scales". He mentioned that can-do statements were interpreted differently by both teachers and students. The differences in interpretation occur in whether the examinee can perform the function or not, but in how well the examinee can perform the function (e.g., I can arrange for a hotel room or taxi ride) quite easily, easily, with some difficulty, with great difficulty or not at all. Ross believed that can-do statements could reduce the accuracy of self-assessment and the construct validity of the can-do scales.

The Positive Validity of Self-assessment

Shrauger and Osberg (1981) reported that self-assessments seemed to predict academic achievement as well as regular assessments. Self-assessment nurtures students' skills by allowing them to learn information and judge their performance in meaningful rather than rote methods (McMillan and Hearn, 2009). LeBlanc and Painchaud (1985) found positive correlations (ranging from .39 to .53 on subtests and .53 on total score) between results on a self-assessment instrument and a standardized English proficiency exam. They found the self-assessment instrument placed students just as well as the formal placement test they had been using, with fewer changes in classes being reported using the self-assessment instrument. Luoma and Tarnanen (2003) confirmed that self-ratings for second language writing were reasonably as good as other standardized tests.

Krausert (1992) established that students were usually able to self-measure speaking and writing, but not reading. Generally, language self-assessment instruments are well received by students and seem to be accurate tools (Oscarson, 1978). Different factors such as types of tasks, test authenticity, self-assessment training, and proficiency level of the students can affect accuracy, and none of these variables appear so problematic that they could not be taken into consideration (Blanche, 1988; Heilenman, 1991; Strong-Krause, 1997). Boud and Falchikov (1989) conducted a meta-analysis study and concluded that self-assessment could be a valuable supplement to formal second language assessment. The authors also noted factors affecting the reliability and validity of self-assessment should be carefully and further investigated.

Wilson (1999) investigated the validity and potential efficacy for self-assessment of speaking proficiency in English as a second language (ESL) according to the ILR scale. The study examined the correlation between scores obtained from a self-assessment instrument, ratings obtained from the ILR-based Language Proficiency Interview (LPI), and ratings obtained from the Test of English for International Communication (TOEIC). The findings of the study indicated clearly that the level and pattern of correlation between TOEIC scores and the self-rating tend to strongly parallel corresponding levels and patterns of correlation between TOEIC scores and the formal LPI rating. The findings obviously seem to suggest the good value of the ILR scale for self-assessment purposes. The findings appear to extend the evidence regarding the validity of the ILR scale. The study also established that the ILR scale is a valid standard for use not only by professional interviewers/raters, but also by educated second language users/learners and educators (p.26).

Implications of Self-assessment

Using Self-assessment in Education

Sufficient evidence shows that self-assessment contributes to student achievement (Hughes et al., 1985; Schunk, 1996; Sparks, 1991), particularly if teachers provide direct instruction in how to self-assess (e.g., J. Ross et al., 2002). There is also evidence that self-assessment contributes to improving student behavior (Henry, 1994; Nelson et al., 1995). Ross et al. (1998) reported that students prefer self-assessment to assessment by teacher alone and the reason cited by students are the additional benefits of self-assessment:

(1) Self-assessment provides a better understanding and a clear idea of what learners should do.

(2) Self-assessment involves students in setting the criteria for the assessment.

(3) Self-assessment appears to be fair because it enables students to include important performance dimensions, such as effort, that are not usually included in their grade.

(4) Self-assessment enables them to communicate information about their performance (e.g., their goals and reasoning) that is not otherwise available to their teacher.

(5) Self-assessment gives them information they can use to improve their work.

LeBlanc and Painchaud (1985) indicated that self-assessment is helpful for adults because adults usually understand the learning situation they will be in; they also speak a first language, thus they understand what is involved in communicating in a language. Learners also suggested three other advantages of using self-assessment:

(1) Less time is involved in completing a self-assessment survey than with traditional tests.

(2) Problems with cheating and test security issues are eliminated.

(3) Self-assessment involves the students more in making decisions about their education, which increases their responsibility for their own learning.

Boud (1995) indicates that there are two important elements that are essential for self-assessment:

(1) Development of knowledge of self-assessment and appropriate standards and criteria for meeting those standards.

(2) Ability to make judgments about whether the work involved does or does not meet the standards.

These elements force learners to participate in critical thinking and evaluation processes. Self-assessment does not only consist of testing and grading an individual's own skills, but also involves an individual in assessing his/her own performance in any given situation. This process requires that learners have the knowledge to evaluate their own work and that they be honest in their evaluation. Sekula, Buttery, and Guyton (1996) agree that self-assessment is based on useful knowledge about the whole self in relation to educational goals. It asks "How am I doing?" "How can I do better?" Students learn to compare and contrast their works with models and against a set of standards and/or criteria (Bourke & Poskitt, 1997). Dickinson (1987) has three reasons for using self-assessment:

(1) Assessment leading towards evaluation is an important educational objective in its own right. Training learners in this is beneficial to learning.

(2) Self-assessment is a necessary part of self-determination.

(3) Self-assessment lessens the assessment burden on the teacher.

Self-assessment Promotes Learning

The main purposes of engaging students in careful self-assessment are to improve learning and achievement, and to encourage academic self-regulation, or the tendency to monitor and manage one's own learning (Pintrich, 2000; Zimmerman & Schunk, 2004). Self-regulation and achievement are very much related: Students who set goals, make flexible plans to meet them, and watch their progress tend to learn more and do better in school than students who do not. Self-assessment is a core aspect of self-regulation because it involves awareness of the goals of a task and checking one's progress toward them. As a result of self-assessment, both self-regulation and achievement can increase (Schunk, 2003).

McMillan and Hearn (2009) stated: "In the current era of standards-based education, student self-assessment stands alone in its promise of improved student motivation and engagement, and learning. When student self-assessment is correctly implemented, it can promote intrinsic motivation, internally controlled effort, a mastery goal orientation, and more meaningful learning (p. 39)." Self-assessment is defined as a process by which students: (1) monitor and assess the quality of their thinking and behavior when learning, and (2) discover plans that enhance their understanding and skills (McMillan & Hearn 2009).

According to Goodrich (1996) there are some conditions that must be met in order to have an efficient self-assessment such as: (1) awareness of the value of self-assessment, (2) access to the criteria on which to base the assessment, (3) a specific task or performance to assess and practice, (4) models of self-assessment, (5) assistance in completing the with self-assessment, (6) a cue regarding when it is appropriate to use self-assessment and provide opportunities to revise and improve task performance. These

conditions appear to be excessive, but student self-assessment is feasible and is happening around the world from elementary schools to adult education (Deakin-Crick, Sebba, Harlen, Guoxing, & Lawson, (2005).

Goodrich's conditions are regularly practiced in DLIFLC classrooms. DLIFLC has adopted the task-based and student-centered instruction as teaching methodologies. These approaches allow students to use the language naturally in performing real-life tasks. Students are given direct instructions to perform specific tasks. Frequent drills are used to improve performance, and so are introduced to the ILR rating criteria. The criteria play a significant role in identifying the errors students are likely to make, in addition to improving their performance. The ILR provides learners with good information about the different tasks they have to perform successfully to reach a certain level of proficiency. Andrade (2001) recommended that explaining the criteria may increase students' knowledge of the standards for a task and help students produce a higher quality of work.

Self-assessment Reduces Classroom Anxiety

The major concern for students in a category IV language is the length of the 63-week course. At the end of the course students get stressed and their learning suffers and their motivation deteriorates. According to De Saint Leger (2009), anxiety was long ago well known as a major obstruction to learning, generally in its effects on learners' self-efficacy. Tremblay and Gardner (1995) described self-efficacy as the learner's perceived expectation to meet the challenge of language learning in relation to specific tasks, and they suggested that anxiety is a ‘‘weakening component of self-efficacy’’ (p. 508). Speaking is the skill that is most associated with foreign language anxiety because students have to perform in public (e.g., Horwitz &Young, 1991).

Anxiety in the second or foreign language is related to insufficiency in listening comprehension, impaired vocabulary learning, lack of word production, and low grades in the language course. (Gardner et al., (1997) and Horwitz (2001) reported that anxious language learners have a tendency to underestimate their abilities. Interestingly, research conducted by Mills, Pajares, and Herron (2006, 2007) suggested that it is students' self-efficacy beliefs, rather than anxiety, that are closely related to academic performance.

The role of teachers is to assist students to develop better self-confidence and thus decrease their language learning anxiety. Tremblay and Gardner (1995) recommended that teachers should encourage students to set suitable objectives because ''individuals with specific and challenging objectives persevere longer at a task than individuals with easy and unclear objectives'' (p. 508). Mills et al. (2006) recommended encouraging students to agree to planning and monitoring strategies in order to promote more practical and positive linguistic behavior. Both goal-setting and practical uses of strategies allow students to attribute achievement or failure to their own level of effort and strategy use, rather than to factors outside their control such as luck or task difficulty. This ability to set goals use strategies helps students acquire a greater sense of success (Graham, 2004).

De Saint Leger (2009) reported that self-assessment appears to be a good activity to assist students to develop appropriate objectives and to monitor their efforts accordingly. From this point of view, self-reflective activities should not be considered the end point of the process, as they are generally defined in self-assessment research, but as practically ongoing, dynamic activities for reflecting at the same time on past experience and possible future performance to determine learning behavior.

This vision of students as active agents in their own learning is also proposed in Socio-Cultural Theory (SCT). Lantolf and Thorne (2006) discussed the critical notion of organization in education. They described organization as the capability of individuals to assign importance and significance to things and events (p. 143). This ability is thought to be influenced by the learner's own chronological path, learning objectives, and developmental stage. The more learners are able to practice organization, the more they become independent, self-regulated learners (Ridley, 2003).

Ushioda (2003) argued that the motivation to learn is not solely located within the individual but is also socially distributed, and that the social unit of the classroom is instrumental in developing and sustaining the motivation of individual learners. The suggestion is that if learners recognize the learning environment as supportive rather than inhibiting, their self-confidence and motivation to interact in the classroom will grow accordingly (De Saint Leger, 2009).

Empirical Research

Immersion and Self-assessment

Since 2005, some of DLIFLC's best performing students have had an opportunity to travel on three- to six-week long, in–country (OCONUS) immersions. During in–country immersions students attend more than thirty hours per week of classroom language and culture study at the host institute and participate in daily out-of-class activities and weekly field trips and tours (DLIFLC, 2011). The immersion training is different from classroom learning, where the learning is facilitated through controlled input and graded on what students can do in the classroom, not in real-life communication tasks. This study investigates whether students who participate in an in-

country immersion are more accurate in assessing their own proficiency than those students did not enjoy that immersion opportunity.

The theoretical theory of immersion programs is based on the assumption that language is acquired through comprehensible input in real-world tasks (Swain, 1985). This is in accordance with Krashen's Comprehensible Input Hypothesis (1985) which states that to be exposed to an environment rich in comprehensible input is adequate for acquisition to take place. According to Krashen, language input that is comprehensible exceeds somewhat the learner's current knowledge of the language (i+1) and is offered in a low-affective-filter environment (Stein, 1999). The idea of immersion training comes from the accepted principle that the only way to learn a language is to go to the country where it is spoken and simply "immerses" oneself in it. Most second language acquisition theorists endorse the input hypothesis in some form. However, many also disagree that comprehensible input is enough for language acquisition to take place (e.g., Swain, 1985; White, 1987).

Cognitivist theories (that is, theories based on the role of active intelligence) have now developed into constructivist theories of learning. Cohen, Manion, and Morrison (2004) explain that "At heart there is a move away from instructing and instructivism and towards constructivism" (p.167). Classrooms are viewed primarily as places in which pupils learn rather than being mainly places in which teachers teach.

Piaget (1953) investigated how the learner increases understanding. Children's minds are not empty, but dynamically process knowledge. Children cannot carry out specific jobs until they are psychologically grown-up enough to do so, and the willingness to learn and progress is different for each individual. There is an emphasis on

discovery learning through practice rather than teacher-transmitted information. Piaget assumed that language develops through interaction with others in real-life tasks.

Vygotsky (1962) recognized that Social Constructivism places more importance upon the role of experience in understanding and meanings developed through collaboration and interaction. The theory of the "Zone of Proximal Development" (ZPD) was developed by Vygotsky. 'Proximal' means next and the ZPD is the gap between a child's current level of progress with no adult help and the level of possible development when working in groups with adults who are competent in the child's language.

It is not important that the adult who is teaching students how to do tasks or functions well; however collaboration and interaction activities develop new understandings of how learners can perform any task or function assigned by an adult. According to Vygotsky, the development of language and the articulation of ideas are central to learning. The learner's current level of speaking reflects the importance of prior influences and knowledge. The learner is stretched, and ZPD is about "can do with help." The teacher's role is to locate learning in the ZPD (Marta, 2007).

Bruner (1960) identified learning as a process of dynamically acquiring knowledge in which learners create new ideas based upon their present and past knowledge. 'Learning how to learn' is central; the development of learning is as significant as the product, and social interaction is important. Bruner proposed three methods of thinking, the inactive, the iconic, and the symbolic, which increasingly overlap each other. Instructors can improve learning by using these three methods. At the inactive level, we can see the significance of the use of physical actions and play, and we can see the managing of real-life tasks. In the iconic method we use pictures or words in

color to present the instructional materials and other objects. Students begin to use the symbolic mode (words and numbers) to think and reason in the abstract and begin to make sense of their experiences as they use the target language in real-world situations.

Wood, et al. (1976) explained scaffolding as the support that teachers provide to assist a student to complete tasks and construct learning that they would not be able to complete based on their own ability. The student starts to develop intellectuality and autonomy when the scaffolding is slowly removed. The scaffolding allows the teacher to increase the student's participation beyond his/her existing abilities and levels of understanding within the ZPD.

Experiential Learning

Experiential learning is based on a practice derived from Piaget and Vygotsky of 'learning by doing' or 'dynamic learning' in which the instructor makes the knowledge to be learned available to the students, who experiment and make discoveries themselves (Marta, 2007). Students learn through their own experience. This is in contrast to systematic language immersion teaching, where there is more focus on the structures of the language. Good practice ensures that both methods ('Experiential' and 'systematic' instruction) are utilized to avoid the threats that arise if one of them is allowed to control the other.

For instance, although experiential language education has a focus on content and use of authentic second language in class, there is a trend for teachers to do much or most of the talking, which can limit the learner's opportunities to actually experience or practice speaking. Systematic immersion education, on the other hand, focuses more on the immersion language structures, but might over-emphasize accuracy to the

disadvantage of communication. This highlights the need to develop classroom strategies that encourage output and intuition, but also structures. Some such strategies are discussed in the next section on Target-Based Language Learning (Marta, 2007).

The Role of Age and Experience in Self-assessment

The research reviewed in this study concerning self-assessment did not indicate that age was a variable in whether students over or underestimate their language skills in comparison to other variables of self-assessment. Mabe and West (1982) point out that the majority of older students usually have more experience in a specific subject. The authors recognize that older students either underrate or rate themselves more accurately. Falchikov and Boud (1989), in their meta-analysis, point out that many studies did not indicate the level of courses from which participants were drawn. However, the authors summarize those finding which did not include this information. The study also points out that "knowledge within a specific field is more influential than duration of the course" (Falchikov & Boud, 1989, p. 425). Stanton (1976) indicated that self-assessment is really not efficient for undergraduates, for the reason that they do not have enough academic experience (p.238). This is a very important point, particularly in light of calls to use self-assessment methodology for such applications as placement testing in foreign language education (Moritz, 1995).

Overall Accuracy of Self-assessment

The most significant question of many researchers was basically whether or not students can assess their own skills or performance. The answer to this question stays unclear, as of the studies reviewed by the researcher report a complete lack of facility on

the part of students to evaluate their abilities. Edwards (1989) points out that the students usually appear to be very critical of their own work. Research on self-assessment in the last twenty years has involved comparing the accuracy of students' self-assessment with so called objective evaluations of individual performance or ability (Moritz, 1995). Unfortunately, it is not feasible to draw any conclusions about the accuracy of self-assessment methods, in large part because of the methodological flaws in most of the research, and inconsistent or incomplete descriptions of empirical situations and results.

Falchikov and Boud (1989) identify problems, such as the contradictory definitions of agreement between teacher and student's ratings, the practice of having students envision their own grades, and the practice of having students include effort in the evaluation of their work. Boud and Falchikov (1989) point out additional concerns in the reported research, for example variables that are incorrectly defined and different rating scales (p.396). The different types of statistics that are used to present the research results add more complication when comparing self-assessment studies (Moritz, 1995). Falchikov and Boud (1989) identify many examples in which results appear to show a tendency in one direction if reported in the form of correlations, which indicate performance relative to the rest of the group. However, if reported in the form of effect size, which simplifies percentages or absolute performance results, subjects will show an opposite tendency (p.426).

Gender Differences in Self-assessment

Looking at gender differences in studies of writing proficiency, Jewell and Malecki (2003) found that girls earned higher scores for writing proficiency than boys in looking at the total number of words spelled correctly within a limited period of time.

When looking at the production and the accuracy of handwriting, gender was not found to be a factor. Jewell and Malecki concluded that boys and girls differ in the amount that they write within a given time limit, but, that their writing accuracy is not significantly different (p. 380).

Pallier (2003) and Lepkowski, (2006) studied male and female self-assessment of performance on a general knowledge test and a visual perception task. The researcher found males were more confident than females on both tasks. A comparison by age was also examined and led the researchers to conclude, "Results indicated that the tendency for men to express higher levels of confidence than women in the accuracy of their work appears to remain constant across the life-span" (p. 265). Pallier concluded in the study that the males performed better than the females due to the confidence which was supported by their scores.

The role of gender in education cannot be dismissed; studies consistently demonstrate that girls are outperforming boys in the area of writing assessments. Still, researchers are emphasizing that educators should not look at these data without questioning the reasoning behind this information and should consider what can be done to enhance male performance in writing proficiency (Mowrer, 2006). This leads to the question of what type of evaluative tool could provide educators with information regarding students' speaking proficiency in an equitable manner for males and females.

Although the majority of self-assessment studies do not describe their subject pool beyond the total number of participants, some of the research that has examined gender differences between subjects is worth study (Moritz, 1995). Bailey & Lazar (1976) compared 20 men and 20 women who completed both a self-rating scale of

college ability and a Concept Mastery Test (CMT). Their results appeared to show that "perceived satisfaction with intellectual ability was a better predictor of actual ability for women than for men" (p.286). They concluded that the accuracy of self-assessment reveals the developmental processes which favor girls, probably due to their greater ego-involvement in achievement and academic abilities (Moritz, 1995).

Swanson and Lease (1990) conducted a study in which they gave 59 women and 53 men a self-assessment questionnaire which they compared with the students' American Comprehension Test (ACT) scores. They also had subjects rate a same-gender colleague on the 30 skills covered by the self-assessment questionnaire. They found that "women and men rated themselves in ways that may match to gender-stereotypic patterns" (Swanson & Lease, p. 351). That is, women rated themselves lower on numerical aptitude, physical dexterity, and mechanical skills. They also found that "participants assessed themselves differently than they did their peers, and that gender differences existed in the peer rating." That is, "women rated women higher, but men rated men lower" (Swanson & Lease, p. 352). Swanson and Lease suggest that these result demonstrate the problem of how little is known about "what internal norm or comparison groups an individual considers in making a self-assessment of skills or abilities" (p.385).

Self-assessment of Knowledge

Self-assessments of knowledge are learners' estimates of how much they know or have learned about a particular domain. Self-assessments offer the potential to reduce the burden of developing tests to determine whether the desired knowledge has been gained as a result of participation in a course or training intervention (Sitzmann, et al. 2010).

Self-assessments of knowledge refer to the evaluations learners make about their current knowledge levels or increases in their knowledge levels in a particular domain. Similar to self-assessing job performance (Campbell & Lee, 1988), when learners evaluate their knowledge, they begin with a cognitive representation of the domain and then judge their current knowledge levels against their representation of that domain. Self-assessments of knowledge are typically measured at the end of a program by asking learners to rate their perceived levels of comprehension (Walczyk & Hall, 1989).

Sitzmann, et al. (2010) provided a much needed review and integration of the literature on self-assessments of knowledge. They use meta-analytic methods to sort through a body of work characterized by mixed findings and competing conclusions to provide convincing evidence that trainees generally have difficulty self-assessing their learning. Based on the authors' findings, Sitzmann et al. suggest a more limited role for self-assessments in evaluation research and practice, but stop short of suggesting that self-assessments should be abandoned as a tool to measure learning.

Despite the potential limitations of self-assessments of knowledge, they are used as an evaluation criterion across a wide range of disciplines, including education, business, communication, psychology, medical education, and foreign language acquisition (e.g., Dobransky & Frymier, 2004; Lim & Morris, 2006). Moreover, self-assessments of knowledge are included in many higher education end-of-semester course evaluations to examine teacher effectiveness. Bell and Ford (2007) argue that there remain several unanswered questions surrounding self-assessments of knowledge that must be addressed before we can reach a more definitive conclusion on the viability of these measures.

Sitzmann, et al. (2010) reported in their research that self-assessments of knowledge are only moderately related to cognitive learning and are strongly related to affective evaluation outcomes. Even in evaluation contexts that optimized the self-assessment–learning relationship (e.g., when learners practiced self-assessing and received feedback on their self assessments), self-assessments had as a strong relationship with motivation as with cognitive learning.

The Role of Military Rank and Service in Self-assessment

The need for investigating the role of military rank "officers and enlisted" and type of services (Army, Air force, Marine or Navy) on how accurately people evaluate their speaking was derived from the location of the study. The study took place at DLIFLC which trains linguists in different military services and has different ranks from private to a major. Shadrick and Shaefer (2007) reported that self-assessment has been used by the U.S. Army as an indication of learners' progress during distance learning courses. The U.S. Army can use self-assessment to check skill acquisition development or the need for further training, and self-assessment can be used post-training to scrutinize for shifts in proficiency or performance. Across the range of self-assessment research, the topic of most debate concerns how accurately people evaluate their own ability or performance.

A study conducted by Fox and Dinur (1988), who examined males participating in military training, found that self-assessments were significantly related to commander and peer ratings. The experimental group was told their reports would be compared with those from other sources; the control group was not. Predictive and convergent validities were examined on three criteria: course success, commander ratings, and peer

evaluations. Self-assessments from both experimental and control group were valid; however, those of the experimental group did not yield consistently higher validities. Findings are discussed in regard to their practical ramifications.

Although the experimental group did not show any statistically significant improvement, predictive and convergent validity were found for course success, commander ratings, and peer ratings. For all but one assessment (i.e., commanders' assessments of efficiency under pressure, $p < .01$), differences between groups were insignificant ($p's > .05$). Another positive finding was that there was less of a halo effect for the self-assessments than for the peer ratings. This literature presents an argument for the capacity to effectively and accurately utilize self-assessment, while other studies contend that self-assessment produces under-estimation.

Breidert (2009) investigated the effect of skill level and item ambiguity on the accuracy of self-assessments made by lieutenants and captains in the U. S. Army. The results indicated that increased skill level resulted in increased accuracy of self-assessments while ambiguity had no effect. Respond to the hypothesis, as items became more ambiguous, both captains and lieutenants self-assessed more accurately.

Summary

This chapter reviewed literature in relation to self-assessment theories (Social Constructivism Theory, Multiple intelligences Theory, and The Social Cognitive Theory) as a foundation for this study. It also discussed literature related to the limited and positive validity of self-assessment and the benefits of using self-assessment and its impact on learning foreign/second languages. The literature reviewed the variables that

likely affect the accuracy of self-assessment: (1) Immersion in the target country, (2) Experiential Learning, (3) Age and Experience, (4) Self-assessment of knowledge, (5) Role of Military Rank and Services, and (6) Gender differences.

In Chapter 3, a discussion of the methodology employed in this study will be described. The research design, data collection methods, participants and sample size and ethical assurances will be described.

CHAPTER THREE

RESEARCH METHODOLOGY AND DESIGN

Introduction and Restatement of the Purpose

The purpose of the study was to develop and validate a self-assessment instrument that can be used to obtain a reliable estimate of the foreign language proficiency of native speakers of English. To realize this purpose, the researcher carried out the following steps: 1) investigate the relationship between two types of measures of oral proficiency: level scores inferred from the self-assessment instrument and ratings obtained from a formal Oral Proficiency Interview (OPI). 2) Investigate the impact of various variables that the literature suggests are likely to affect the validity, reliability and accuracy of self-assessment scores and how well students assess their speaking proficiency in a second language or a foreign language.

Research Method

A correlational research method was employed to examine the validity and reliability of self–assessment of speaking proficiency in Arabic, Chinese and Korean as foreign languages, as measured by the self-assessment instrument that the researcher developed, with the OPI as the formal criterion test.

Two research questions were explored in this study:

RQ1: What is the relationship between student scores on a self-assessment survey and those on a criterion- referenced speaking proficiency test?

RQ2: What is the impact of various variables on the psychometric properties (validity, reliability and accuracy) of a student self-assessment of second language speaking proficiency?

Research Design

A descriptive research design was used to collect data from students who had completed 63-week basic courses in three Category-IV languages (Arabic, Chinese and Korean) at the Defense Language Institute Foreign language Center (DLIFLC) in Monterey, California. The primary data collection method was through a self-assessment questionnaire. According to Burns (1999), a quantitative study can be used when the relationships among variables can be identified objectively and the research variables can be measured and/or controlled; thus, this method is suitable for this study which examines which variables may be related to adult learners' assessment of their language proficiency and their scores on OPI tests.

Participants

This study was conducted at the Defense Language Institute Foreign Language Center in Monterey, California. The research population consisted of students in four military services who are studying Arabic, Chinese or Korean as a foreign language. The U.S. government assigns foreign languages to four categories according to their difficulty for native speakers of English. The three languages included in this study fall into Category IV, the most difficult languages for native speakers of English to learn. Participants ranged in age from 18 to 40, and in military rank from private to major. Of the 350 students who voluntarily participated in the study, 83 were females and 267 were males, and the majority were native speakers of English. The number of participants who were not native speakers of English was so small that it was not considered a threat to the validity of the study since they learned English at an early age.

DLIFLC does not have a standard school year because classes are scheduled to respond to customer–agency requirements. Therefore, classes begin and end at different times, and graduations occur on a continual basis throughout the calendar year. The timing of the selection of students was based on the time of graduation in each language program. For this reason, the administration of the instrument took six months.

At the end of the language program, some participants were sent on a four- to six-week immersion abroad to study and practice the language in the target country. The immersion program is training that is provided in the target-language country, where the learning takes place in a real-life environment. Students believe that the best way to learn a language is to go to the country where it is spoken and simply "immerse" oneself in it. DLIFLC has developed study abroad programs where students learn and practice the language in the target country; however, these immersion programs were not available to all students in this study due to financial issues or political unrest in the target countries.

The length of the basic-course program for Category IV languages is 63 weeks. Students receive six to seven hours of language instruction a day. The military services' expectation is that students should score between level 1+ and 2 on the speaking test at the end of the program. Less than one percent of students score level 1, and less than two percent of the students score at level 2+.

Instrumentation

Can-Do-Scale (CDS) and Oral Proficiency Interview (OPI)

The CDS was adapted from a self-assessment of foreign language speaking proficiency on the ILR website. Using this source instrument in the study did not require prior permission (Parry, 2011). The federal government developed the ILR criteria. The

ILR criteria and the original version of the self-assessment are in the public domain at the ILR website. The original version of the CDS had not undergone any previous validation or reliability test. As a result, a validation process was conducted to ensure that the native speakers of English who participated in the study would have no difficulty comprehending the can-do items, because ambiguously worded items in an instrument pose a threat to the instrument's reliability and validity. Burns (1999) reported that in developing can-do items much attention must be given to students' language level, the clarity and brevity of the questions, and the extent to which learners have the knowledge required to answer the questions.

Participants in the study were asked to respond to 30 "can-do" statements related to how well students can perform tasks in real-life situations. The instrument was designed to measure foreign language speaking ability from Level 0 to 3, including plus levels. The survey was constructed with a bottom-up approach; it started with the lower-level items and then continued up to level three.

Development and Validation of the CDS Instrument

The original self-assessment instrument for speaking proficiency at the ILR website consists of 39 "can-do" statements that cover five levels of proficiency from level 1 to 5. The instrument is based on a two-point scale. Participants have to answer each can-do statement yes or no. A "yes" response means the participant consistently performs the task or function described successfully. If the participant thinks the statement describe his/her ability only some of the time, or only in some contexts, he/she should answer no. In modifying the original instrument, the researcher expanded the scale

to five points to increase the learners' opportunity to provide a more fine-grained measure of their proficiency.

The participants in this study are basic-course students, and because the speaking scores of such students almost never reach level three on the ILR scale, the instrument was reduced to 26 statements covering three levels of proficiency from level one to three. The instrument was then expanded by adding four items taken from a DLIFLC guide, *"A Guide for Evaluating Foreign Language Immersion Training"* (PRC, INC. 1997). This guide contains a compendium of can-do statements that can be used to evaluate immersion training. Permission was granted to use four of the can-do statements included in this publication (see Appendix B). The final version of the instrument used in this study thus has a total of 30 can-do statements.

The first section of the instrument included questions about student background, including, name, age, gender, military rank, military branch, target language, academic background, experience in second language(s), and immersion in the target country. The second section employed a five-point Likert scale to elicit information about students' self-assessment of their ability to speak the target language. In this second section 30 can-do items representing three different proficiency levels are presented sequentially from level one to level three.

The self-assessment questionnaire went through an intensive validation process. The original version of the instrument is directly tied to the ILR scale and employs testing jargon that might not be familiar to the average native speaker of English. The ILR terminology and phrases were simplified to the point that they do not generate misunderstanding or different interpretations.

The study was introduced at a staff meeting of the Evaluation and Standardization Division of the DLIFLC. A panel of five ILR experts, including a former Vice Chancellor of Evaluation and Standardization, assisted in the validation.

They were:

Dr. Martha Herzog: Former Vice Chancellor of Evaluation and Standardization

Dr. John Lett: Dean of the Research and Analysis Division

Dr. Monika Ihlenfeld: Dean of the Proficiency Standards Division

Dr. Mika Hoffman: Dean of Test Development

Dr. Gordon Jackson: Research Specialist

Mr. James Dirgin: Director of Test Review Education

The validation process went through several steps. First, the researcher scheduled meetings with each member of the group to describe the study and to explain what was expected of him/her as a member of the validation panel, and the steps in the validation process. After meeting with the entire panel, the researcher emailed the original self-assessment questionnaire to the panel members.

Second, the panel members recommended expanding the response scale to a five-point scale instead of yes or no. For example, for can-do statements like "I can order a meal in a restaurant," students select among five alternative self-assessments: (1) quite easily; (2) easily; (3) with some difficulty; (4) with great difficulty; (5) not at all.

Third, the panel made explicit and detailed comments and gave feedback for each item in the survey. The reviewers' comments, concerns and suggestions addressed content, lexicon and syntax. The necessary editing and revisions were made in accordance with the reviewers' suggestions and recommendations.

Fourth, reviewers were given a revised version to review. Reviewers raised more issues in the second review concerning grammar and syntax, but they made fewer comments than in the first review. Finally, the panel members were given another revised version for a final review. The final version of the instrument was approved by the panel and was ready for the next step of testing its reliability through "a test-retest analysis."

Rules for Scoring the Self-assessment

A focus group of master testers was formed in order to develop the rating protocol and the scoring rules for the CDS. A meeting was scheduled with everyone involved in this process to discuss the rating process and the scoring protocol. The group developed the rules on how to score the CDS and the quality control processes. Each member of the group received a hard copy of the rating protocol. Every member of the group attended a practice rating session on actual samples to norm on the rating protocol to ensure high inter-rater reliability among testers. The survey was scored by two master testers. Third raters resolved any discrepancies between the initial raters.

The ILR Skill Level Descriptions describe six "base levels" (0 to 5) and five "plus levels" (0+ to 4+). A plus is assigned when an examinee's proficiency *substantially exceeds* the proficiency associated with one base skill level but does not fully meet the criteria for the next base level. However, since the CDS covers only Levels 1 through 3, the only plus scores it can generate are 0+, 1+ and 2+. The "plus" levels do not represent thresholds, but describe language that substantially exceeds the criteria of one base level yet does not fully meet the requirements of the next higher base level. Another way of looking at the "plus" levels is that they contain an indication of the next higher level, but not sufficiently sustained to earn the rating of that next higher level.

The survey was scored manually according to the following scoring rules:

1. Students must meet the requirements for the lower base level before looking at the responses for the next higher level. Scorers will therefore begin by scoring the Level 1, can-do items.
2. For each can-do statement, responses of “quite easily and easily” will be treated the same in terms of scoring, meaning that the student is at the level associated with the stated task.
3. Students who answer all can-do statements for any level “quite easily or easily” will be beyond the threshold for that level and will be given credit for being at that level.
4. “With great difficulty and not at all” will be treated the same in terms of scoring, indicating that a student is below the base level associated with the item by a full level.
5. “With some difficulty” indicates that the student is at the plus level for the task/function in view, if the student has already established that he/she meets the requirements of the next lower base level. The student may be at the plus level overall, but that determination can only be made in the context of responses to other items.
6. The CDS has 30 items, 7 items at level 1, 14 items at level 2, and 9 items at level 3.
7. The following chart simplifies the scoring rules (SR).

Table 1: Survey Scoring Rules

SR	L 0+	L 1	L 1+	L 2	L 2+	L 3
R1	7 questions at L1 answered "great difficulty"	2 or more questions at L1 answered "easily" or "quite easily"	3 or more questions at L2 answered "some difficulty"	12 or more questions at L2 answered "easily" or "quite easily"	1 or more questions at L3 answered "some difficulty"	All 9 questions at L3 answered "easily" or "quite easily"
R2		5 or less questions at L1 answered with "some difficulty"	3 questions or less at L2 answered "great difficulty" or "not at all"	2 or less questions at level 2 answered "some difficulty"		
R3		4 or more questions at L2 answered "great difficulty" or "not at all"		1 or more questions at L3 or answered "great difficulty" or "not at all"		

Reliability of the CDS Instrument

The reliability of the survey was established through a test-retest study with alternative forms administered from one to two weeks apart. A group of eighty students studying Arabic participated in the test-retest study. The group consisted of forty students in semester I, sixteen in semester II, and twenty four in semester III. The self-assessment instrument was scored independently by two certified and experienced master testers according to the scoring protocol set by researcher.

Each of the master testers, working independently, scored forms A and B of the CDS for his/her language group. The forms (A & B) of the CDS were developed from the same set of items by presenting the items in different order. The items' orders were scrambled within each proficiency level. Participants were asked not to write their names on the printed CDS for two reasons**:** (1) to minimize any rating bias among testers. (2) To protect human subjects' confidentiality.

Each student was given a code from a master list. Students wrote the code number instead of their names on the two forms of the instrument. Testers scored both forms globally to check the reliability of the parallel forms of the instrument. A spearman's rho correlation coefficient was used to compute a correlation between the scores on the two forms. The results showed that the test-retest of the semester I students was statistically significant ($r = .638$, $p < .05$) (see Table 4). In semester II the results showed the test was statistically significant. ($r = .902$, $p < .05$) (see Table 5). In semester III the results showed the test was statistically significant. ($r = .803$, $p < .05$) (see Table 6).

When two variables are measure by two individuals to measure the same thing, one can use Cohen's Kappa (often simply called Kappa) as a measure of agreement between the two individuals. Cohen's Kappa was also used as a measure of test-retest reliability. For semester I was $k = .566$, and semester II was $k = .719$, and semester III was $k = .574$.

In case there were discrepancies in the scores between the two testers, another master tester would score the survey to resolve the discrepancy and ensure scoring reliability. The testers assigned a numerical score for each self-assessment questionnaire corresponding to each proficiency level. These are the level-codes used in the survey

(CDS), (Level 0+ = 6) (level 1 = 10), (level 1+ = 16), (level 2 = 20); (level 2+ = 26), and (level 3 = 30).

Administration of the Survey

A formal meeting was arranged with the deans of five schools (three Arabic, one Chinese and one Korean) to request their cooperation and explain the procedures of the study. Administration of the self-assessment questionnaire took place over a five-month period based on the time of students' graduation. The questionnaire was administered to students one to two weeks before the final OPI test to minimize the likelihood that any learning that took place during that one-to-two week period would affect the OPI score.

Administration of the self-assessment took place during the school day in order to reduce any disturbance of the students' military activities. The graduation timetable was discussed and confirmed with each dean to ensure that the days and time selected had no conflict with any planned activities by the schools. Each school was reminded three weeks in advance of the date scheduled to ensure there were no changes in the schedule. The researcher introduced the study to the participants to solicit volunteers.

The presentations were identical every time he met with the students before they took their end-of-course OPI test. Students were informed about the purpose of the study, its risks and benefits, as well as data protection and confidentiality. Students were informed that the researcher was the only person allowed to see their surveys and that all personal information would be kept highly confidential. Students who volunteered to participate in the study signed the informed-consent form (see Appendix D) before completing the survey. Students were instructed to complete the survey honestly and to the best of their ability. The self-assessment questionnaire was administered in a paper-

and-pencil format. The time spent completing the self-assessment ranged from fifteen to twenty minutes.

Oral Proficiency Interview (OPI)

The Role of the OPI

The OPI *evaluates general proficiency.* General proficiency is the capability to carry out real-world communication tasks, including every day conversation and work-related tasks pertinent to the individual, either within the target culture or during encounters with individuals who are native speakers of the target language. Examinees with higher proficiency levels can complete tasks of increasing difficulty. Proficiency is unrelated to how or where the examinee acquired the language. The OPI is a holistic measurement method because it evaluates language production in a global, overall manner by determining patterns of strengths and weaknesses, establishing a speaker's level of sustained, practical ability (floor) as well as the clear upper limitations of that ability (ceiling). It does not evaluate discrete features of language use or knowledge about the language. The full descriptions of the ILR contain information about which language tasks the examinee is able to carry out at each level, as well as the quality of the performance.

OPI – Proficiency

The principle of language proficiency testing is to measure how well the examinee employs the language in real-life situations. In contrast to achievement testing, proficiency testing focuses on overall language skill without any consideration of the place, length of time, or method in which that skill was acquired. The OPI evaluates

spoken language by comparing a person's performance of specific language tasks with the ILR scale.

Since a proficiency test is not based on a particular curriculum, it is not possible for an examinee to anticipate what specific questions will be asked. In a proficiency test, examinees will always be asked questions for which they have not prepared. Testers systematically sample the examinee's language and seek to elicit the highest level that the examinee can sustain. Some questions that may be difficult for the examinee are posed in order to find the upper limits of his/her ability. In this way the OPI allows testers to assign a global rating that illustrates what a speaker can do with the language.

OPI Characteristics

The test is a structured, task-based conversation between the testers and the examinee. Within this structure, each test is unique, with the content/topics reflecting the examinee's background, life experiences, and interests. Each question selected by a tester, to elicit language from the examinee, is generally determined by the examinee's prior responses; and the level of difficulty of questions is adjusted continuously according to the examinee's performance. For this reason, conducting an OPI is an interactive and adaptive process.

The specific content of the OPI is determined in large part through conversation, depending on information offered in response to the tasks posed. There are prescribed tasks that the tester must pose at each proficiency level. A tester must elicit OPI tasks by asking questions based on a continuous assessment of the examinee's ability, in the topic areas that are mentioned in the "warm-up" or emerge later in the conversation.

OPI Validity

According to Bachman (1990), validity generally means measuring only the intended knowledge, skill, or ability that one wants to measure. There are different types of validity, such as content validity, face validity, and construct validity. The OPI has a high degree of face validity because it tests speaking by having people speak. Face validity of the OPI as seen by the interviewee is often a judgment call on the part of the interviewee as to how well a test functions for a given purpose.

Content validity in the OPI is dependent upon whether the test contains the tasks and context/content areas characteristic of spoken language skills at a particular level, and includes enough probes (more difficult tasks) to prove that the examinee is not at a higher level. Assessing a learner's proficiency means testing his/her ability to function in the target language (TL) in simulated real-life situations. For example, the test content and questions for a beginner who knows only the class material covered in a few weeks are different from those used with a learner at proficiency level 2 on the ILR scale.

OPI Reliability

According to Bachman (1990), reliability refers to the consistency of results, and these are approaches to establishing the reliability of an instrument. Although there is, strictly speaking, only one type of reliability, there are several ways of establishing reliability. The test-retest approach typically involves the use of "alternate" or "parallel" forms. To accomplish high inter-rater reliability, training is indispensable for OPI testers. An examinee's performance on a test could be affected by factors such as fatigue, weather, time of day, and other external variables that are not being tested. These factors

could contribute to different test results for the same person if s/he could take the same test again, under different circumstances.

Reliability through retesting is established by testing the same examinee twice within a short period of time, during which little or no learning is assumed to have taken place. In this test/retest situation, scores should fall within an acceptable range. For example, if a student scores 88 on a test on Tuesday, and scores 90 on an alternate form on Wednesday, then the test may be said to be reliable. The scores are not absolutely the same, but they fall within an acceptable range. However, if the student had gotten 75 on the first test and 97 on the second administration of that test, the test could be regarded as unreliable if the same tendency were seen with other students.

Reliability through parallel forms refers to the degree to which alternate (parallel) forms of the same test generate equivalent results. This approach offers an approximate means of estimating a test's reliability when internal consistency is either an inappropriate criterion or is simply impossible. Using parallel forms of the same test is particularly useful in minimizing security problems that arise when the test cannot be given to everyone at the same time (as is the case in OPI testing). Also, as in some language maintenance programs, it might be necessary to give the same test to the same person repeatedly over a period of time, and the reliability of that test would ultimately lessen due to the "practice effect." The use of parallel forms eliminates the problem of the practice effect.

The proper administration of the OPI uses the principle of parallel forms to good effect: Although the content areas (that is, the specific topic selection and the details within each topic) vary from one OPI to the next, based on the unique interaction

between each individual examinee and his/her two testers, the same format and test structure are imposed on each OPI. This makes each OPI a parallel form of every other. Thus, one examinee could be tested twice in a period of a few hours, and the content of the two OPIs could be completely different. This feature of the OPI minimizes both security risks and the practice effect.

Inter-rater reliability refers to the degree to which two or more testers independently assign the same rating to the same examinee. Inter-rater reliability is checked by comparing the scores given by one tester to those given by his/her partner. Ideally, in the OPI, the scores should be highly correlated or be within a plus point of one another, and any subsequent ratings of the same test by other raters may also fall within that range. As indicated above, inter-rater reliability can also be computed for a whole group of testers. A high degree of agreement among testers indicates that they understand and apply the scale in the same way. OPI testers are given extensive training and are subjected to random quality checks to assure that they are normed to the ILR standards.

Reliability of OPI Testing at DLIFLC

All OPI testers at the DLIFLC are recertified annually to maximize reliability. The inter-rater reliability of OPI testers in Arabic, Chinese, and Korean for 2011, as measured by the percentage of tests in which the original, independent ratings of the two testers were identical, was 95.89% in Arabic, 98.36% in Chinese, and 95.89% in Korean. The above data were retrieved from the DLIFLC's Testing Division Database in March 2012.

When testers rate an interviewee's performance, they rate independently and each tester puts his/her scoring sheet in a closed envelope no later than 15 minutes after conducting the interview.

Regarding quality control, the test administrators schedule master testers who have extensive experience in testing and receive more training than regular testers to perform this function. A minimum of twenty percent of the recorded OPIs for each graduating class have to be randomly checked for quality control. Quality checks are also performed in the following cases: (1) when the two testers do not assign the same rating, which leads to a discrepancy in the score; (2) students score below the graduation requirements; (3) students score level 3 in reading and listening comprehension and score level 1+ in speaking; (4) there are outlier scores (scores that lie outside the expected range for graduates of a basic course): level 1, unusually low, or level 2+ unusually high. All of these cases above raise the percentage of the interviews subjected to quality control to approximately 45% in each graduating class.

The OPI tests in this study were conducted face to face. All the scores were entered in the DLIFLC's Proficiency Standards Division (PSD) database within a week after the OPI was conducted and quality control was performed.

OPI Practicality

Practicality refers to cost-effectiveness, to the ease of scoring and administering a test. It is important to consider these factors when deciding to use a test. Like many other endeavors, testing is a question of time and money. The OPI offers a practical way to assess speaking proficiency. An OPI can be completed in twenty to forty-five minutes, depending on the level of the speech sample elicited. Two certified testers elicit a speech

sample by conducting a friendly, yet highly structured, conversation. A series of judgments must be made to arrive at a rating based on clearly defined criteria that allow for efficient scoring. By following strict elicitation and rating procedures, the testers can complete the process in twenty to forty-five minutes. The practicality of the OPI is limited to testing DLIFLC students and personnel from other federal agencies. Therefore, the purpose of this study is to validate a self-assessment instrument that can be used to test thousands of military linguists in the field who are located in U.S. and abroad. Also, the CDS can be used as a diagnostic assessment tool to measure students' strengths and weaknesses and as an alternative assessment instead of the OPI for placement purposes for students who are going to enroll in refresher, intermediate or advanced courses.

OPI Content and Context

The Oral Proficiency Interview (OPI) is a test of general language ability, not the ability to converse in the special language or jargon of a particular field (e.g., law, engineering, medicine, etc.). Content is the most variable of all four sections of the assessment criteria used in rating an OPI, task/function, content/context, accuracy, and task type. Content shifts according to topic, setting where the test is given, and educational level of the examinee(s). Tasks refer to a list of global, or overall, language functions that examinees must be able to perform to qualify for a rating at a specific level. Accuracy refers to how well the examinee uses grammar, lexicon, pronunciation, fluency, sociocultural appropriateness, and discourse to perform tasks or functions.

Testers can find that content varies widely from one examinee to another. The content areas addressed in an OPI are more limited at the lower proficiency levels, where the OPI may more closely resemble an achievement than a proficiency test due to the

limited nature of the topics the examinee can address. At lower levels, too, the tester's basic responsibility is to discover an area of interest that the examinee is comfortable talking about, and then find other areas to probe the examinee at the next higher to establish "the ceiling of the test" or what the examinee cannot do with the language and failed to do the functions. At higher levels the content areas that an examinee can handle broaden considerably.

OPI Accuracy

Language accuracy refers to the acceptability, correctness, quality, and appropriateness of the message conveyed. Accuracy can also be observed in the areas of fluency, pronunciation, lexical control, structural control, and social-cultural appropriateness. As in the areas of task/function and content/context, accuracy requirements increase with each level as one ascends the scale. Accuracy plays a key role in the test, particularly when the examinee is speaking at level 2 or higher. One of the criteria for level 2 is the ability to make oneself understood to native speakers who are not accustomed to speaking with non-natives. To achieve this, the speaker must be able to speak without making too many errors that might interfere with communication.

The OPI as a Criterion-Referenced Test

The OPI is a criterion-referenced, rather than a norm-referenced, test. Each language sample is measured solely according to the criteria of the rating scale rather than being compared to the performances of other speakers. Because of the global, holistic nature of the test procedure, which takes into account four assessment factors– task/functions, context/content, accuracy, and text type–a variety of individual

performances will fall within the same proficiency range. Nonetheless, each individual performance must minimally meet the criteria in all four factors for a given level, in order to be assigned to that level.

Categories of Assessment Criteria for the OPI

Although performance on the OPI is globally measured, there are four main categories of assessment on which ratings are based. They are: (1) the global tasks and functions performed with the language (e.g., asking and answering simple questions, narrating in the past, present and future, describing people or things, supporting an opinion about societal issues, and tailoring the language from formal to informal); (2) the social contexts and the content areas in which the language can be used (e.g., formal or informal settings in which these tasks are performed and topics that relate to these settings); (3) the accuracy features that define how well the speaker accomplishes the tasks pertinent to those contexts and content areas (e.g., structure, lexicon, delivery (pronunciation, fluency, and sociolinguistic appropriateness, i.e., "the acceptability of what is being said within a certain setting"), and the use of appropriate discourse strategies; (4) the oral text types produced (e.g., discrete words and phrases, sentences, paragraphs, or extensive discourse).

Procedures of the Study

The study included three languages, Arabic, Chinese, and Korean, which are being taught in five schools at the DLIFLC. Samples of 220 Basic-Course students were recruited in Arabic, 65 students in Chinese, and 65 students in Korean. The selection in each language was based on graduation date and the time when the data were collected.

Each language school consists of five departments and each department consists of small teams of teachers who teach small classes of students who graduate at different times, depending on the starting date of their 63-week course. Multiple visits were made to each school to collect the research data until the sample size needed for this study was reached. The sample consisted of students who were close to graduation.

Participation in the study was on voluntarily basis, and any participants could withdraw from the study at any time without negative consequences. The school deans were contacted to obtain their support and to introduce the study. A visit was made to each team involved in the study to administer the CDS one to two weeks before graduation for about half an hour in coordination with the department chair and the team leaders. The participants were gathered in a room to be briefed on the study and to sign a consent form (see Appendix D) if they decided to participate in the study. Those who volunteered to participate were informed that they could withdraw from the study at any time with no negative impact.

OPI Procedures

The final OPI test took place approximately from one to two weeks after students completed the survey. The purpose of administering the survey shortly before the final OPI test was to minimize the possibility that any learning that occurred after administering the survey would affect the final OPI score. Students were informed about the date and time of their OPI tests approximately two weeks in advance. The testers were selected by means of a computer program designed to ensure that all testers of a given language test about the same number of times per year. The testers selected were notified through their supervisors by the Test Management Division (TMD) two to three

weeks prior to the test date. Each tester was scheduled to conduct two to three tests either in the morning or in the afternoon. A third tester was assigned as a substitute in case one of the two testers could not conduct the test for any reason.

OPI Application and Structures

The four mandatory phases of the OPI are: (1) the warm-up, (2) the level checks, (3) the probes, (4) and the wind-down. If any of these four phases is not there, the speech sample is not ratable. Each phase has a specific function, which can be viewed from three different perspectives: the psychological, the linguistic, and the evaluative. The three perspectives shed light on the significant roles in, or aspects of, the OPI: the speaker/examinee; the speech "sample" or language produced; and the testers who evaluate the examinee's language.

The psychological perspective considers how the examinee experiences speaking the target language, including the conditions that can affect the examinee's performance in the test. The linguistic perspective is interested in the content and quality of the speech produced in the performance of specific tasks in the language. Finally, the evaluative perspective regards the language produced during the tasks as evidence of the examinee's general proficiency in the target language. The speech sample is measured against the criteria laid out for each level in the ILR. In the following introduction to the four phases of the OPI, we will consider the relevance of the three perspectives to each phase.

Phase 1: Warm-Up

Every OPI begins with a warm-up. Testers should begin speaking to the examinee at a normal rate of speech, exchange greetings, and initiate a friendly conversation.

During the warm-up, the tester should not try to challenge the examinee in order to maintain his/her self-confidence. Rather, the warm-up allows the examinee to ease into speaking the target language. Regardless of their proficiency level, some examinees will need time to adjust to using the target language, especially if they do not speak it often. The warm-up also gives the testers a chance to establish a professional and friendly rapport with the examinee and to set the tone of the test. The warm-up should reflect the language and culture being tested. For instance, what is appropriate in American English may not be appropriate in Japanese. From a psychological point of view, the purpose of the warm-up is to make examinees comfortable. From a linguistic point of view, it allows the examinees to adjust to speaking the language and to get used to the testers' pronunciation and way of speaking. Finally, from an evaluative perspective, it provides a preliminary indication of the "working level" (see below) where the examinee shows the highest accuracy. This initial phase is also a crucial first opportunity for the tester to make a mental note of topics that can be developed during subsequent phases of the OPI.

The warm-up should take only 5-7 minutes, yet it is an important part of every test. The warm-up takes the form of a three-way informal conversation. This means that testers may address a topic already introduced by a co-tester to obtain more information.

Phase 2: Level Checks

Each proficiency level has specific types of speaking tasks associated with it. The first question or speaking task that the testers give the examinee after the warm-up is at the proficiency level they estimate the examinee has, based on his/her performance in the warm-up. This estimated proficiency level is referred to as the level checks or the working level and it may be raised or lowered during the interview.

Level checks have the psychological purpose of providing opportunities for examinees to see for themselves, and demonstrate to the testers, what they can do with the language. From a linguistic perspective, level checks identify the functions and content areas that examinees can or cannot handle in accordance with the ILR level descriptions. From an evaluative point of view, level checks establish the "floor" (the ILR level at which examinees are able to sustain their performance.)

The working level is the ILR level that the testers consider to be the actual proficiency level of the examinee. It is based initially on the evidence demonstrated during the warm-up, but may be raised or lowered during the test, depending on the examinee's performance during various level checks and probes. During the OPI, testers must continually verify the working level, changing it as necessary. By the time the testers are ready to begin the wind-down, the working level should be the examinee's score, subject to final verification through the formal rating process.

Phase 3: Probes

Probes are the testers' attempts to raise the level of the examinee's language by posing tasks at the next higher "base level" (OPI, 2000). From a psychological point of view, failed probes indicate to a tester what the examinee cannot do with the language. From a linguistic point of view, failed probes identify the tasks and content areas the examinee cannot handle, resulting in "linguistic breakdown." From an evaluative point of view, there are three possible results of probing :(1) the examinee cannot handle any of the linguistic tasks presented in the probes, confirming the level established during the level checks; (2) the examinee handles each probe consistently and accurately, thereby raising the working level–that is, establishing a new floor at the next-higher base level.

(Note: successful probes count as level checks at the new, higher working level.); (3) the examinee neither completely fails nor altogether succeeds in performing the linguistic tasks presented by the probes. The degree to which the performance at the next higher level is not sustained will determine whether or not the testers assign a plus rating.

Phase 4: Wind-Down

The wind-down serves the psychological purpose of ending the test on a comfortable level and thus finishing the OPI on a positive note. From a linguistic point of view, it gives the tester the opportunity to ensure that the test process is complete. Unlike the other OPI phases, however, the wind-down has no significance from the evaluative perspective. The wind-down does not add anything to the speech sample, and it should not be taken into account in the final rating.

DATA ANALYSIS

This study employed two questions to investigate the relationship between students' self-assessment and a formal assessment in the form of the Oral Proficiency Interview (OPI). The research questions and their associated statistical hypothesizes are as follows:

RQ1: What is the relationship between student scores on a self-assessment survey and those on a criterion-referenced speaking proficiency test?

H1: There is a relationship between student scores on a self-assessment survey and those on a criterion-referenced speaking proficiency test.

$H1_0$: There is no relationship between student scores on a self-assessment survey and those on a criterion-referenced speaking proficiency test.

RQ2: What is the impact of various variables on the psychometric properties (validity, reliability and accuracy) of student self-assessment of second language speaking proficiency?

H2: There is an impact of various variables on the psychometric properties (validity, reliability and accuracy) of student self-assessment of second language speaking proficiency.

$H2_0$**:** There is no impact of various variables on the psychometric properties (validity, reliability and accuracy) of student self-assessment of second language speaking proficiency.

Research question 1 investigated the relationship between students' self-assessment and the formal assessment of OPI scores. Percent agreement was calculated to show how often examinees received the same rating on the CDS and OPI. A correlational approach was also employed to examine the relationship between the CDS and OPI scores. The advantage of using the correlational method was the ability to demonstrate a positive or negative correlation between the CDS and OPI scores.

Spearman's rho was used to calculate the correlation between the two scores. Because CDS and OPI scores represent ordinal rather than interval data, they were converted to ranks. That is, the ILR scores of the OPI and CDS (e.g., 0, 0+, 1, 1+ etc.) were converted to ranks (0=1, 0+=2, 1=3 etc.). A bivariate correlation procedure in SPSS (Spearman's *rho* correlation coefficient for rank orders) was employed to examine the relationship between the two measures of CDS and OPI.

An analysis of the percent agreement between the two forms of assessment was utilized to determine the percentage of participants who got the exact same score on

both measures. This is called “perfect agreement.” For example, the examinee received 1 for each paired rank scores that had an exact agreement and 0 if there was a disagreement between both scores.

Research question 2 investigated the impact of various variables on the psychometric properties (validity, reliability and accuracy) of student self-assessment of second language speaking proficiency. Logistic regression approaches were used to answer this second question and to determine whether each of a set of independent variables had a unique predictive relationship to a dichotomous dependent variable, which in this case was perfect agreement between the CDS and OPI ratings.

Separate logistic regression analyses were conducted in SPSS for each of two dependent variables the first was Perfect Agreement (0 or 1) as a categorical dependent variable. The independent (categorical) variables of (1) Immersion in Target Country, (2) Gender, (3) Age, (4) Education Level, (5) Military Rank, (6) Military Branch, (7) Prior Experience in Studying Foreign language(s), (8) Type of Foreign Language, (9) Heritage Speaker, and (10) Students Who Qualified at ILR Level 2 or Higher were examined to determine which variable(s) had an association in predicting the level of agreement. .

The second dependable variable was called “Within Range.” It showed how far each participant’s CDS score was from his/her OPI score. If the scores of both measures varied by +/- a plus level, they were said to be “Within Range.” However, if the scores of both measures varied by +/- a whole level they were “Outside the Range.” For example, an examinee who had a CDS score of L2 and an OPI score of L1+ was “within range of the target” or only “slightly off target.” Alternatively, an examinee who had a CDS

score of L3 and an OPI score of L2 was “Outside the Range”, or “far off target score by a whole level.”

The advantages of using the logistic regression analyses are: 1) it is more robust than linear discriminant analysis, 2) the independent variables don't have to be normally distributed, or have equal variance in each group; 3) it does not assume a linear relationship between the independent variables (IV) and the dependent variables (DV); 4) it can handle nonlinear effects; 5) you can add explicit interaction and power terms; 6) Normally distributed error terms are not assumed; 7) it does not require that the independent variables be interval.

Unfortunately, the advantages of logistic regression come at a cost: it requires much more data to achieve stable, meaningful results. With standard multiple regression, and Discriminant Analysis (DA), typically 20 data points per predictor is considered the lower bound. For logistic regression, at least 50 data points per predictor are necessary to achieve stable results.

Location of the Study

This study took place at the Defense Language Institute in Monterey, California. Research data were collected from students in three language programs: Arabic, Chinese, and Korean, which were taught in five separate schools. Each school is headed by a civilian dean, who is in charge of planning and implementing assigned programs in foreign language education, curriculum development, implementing academic and administrative policy, and managing the school’s annual manpower and budget.

An associate dean, who is a senior military officer, provides counsel and assistance to the dean, monitors student progress, and directs the school’s Military

Language Instructor (MLI) Program. MLIs are provided by the military units to each school and serve as liaisons between students, faculty, staff, and the units. Aside from performing administrative duties, the MLIs also teach in the classroom and are an essential element in successful language instruction.

Each school is composed of departments, in which instruction in individual foreign languages (FL) takes place. Each department is headed by a civilian chairperson, who is responsible for the instructional program, manages instructors and staff, and supervises foreign language education and the faculty development process. Teachers, organized into teams, are responsible for teaching classes, evaluating student performance, and developing and maintaining course materials.

Protection of Human Subjects (IRB)

According to Flinders (1992), confidentiality protects research informants from stress, embarrassment, or unwanted publicity and also prevents others from using the information participants have provided against these participants. The confidentiality of the students was protected by providing each participant with a pseudo–code which was used in all data discussion and representations, including transcripts. For students studying Arabic the number began with the language code followed by a two to three digits, for example, for student John Smith, the code would be AD01. All documents were digitized and saved on a thumb drive, and were put in a safe deposit box.

CHAPTER 4

PRESENTATION AND ANALYSIS OF DATA

Chapter Overview

This chapter presents the results of the research in a descriptive format including tables and graphs. Results are divided into five sections: (a) pilot study, (b) demographic findings, (c) investigation of assumptions as they relate to inferential analysis, (d) tests of hypotheses, and (e) analysis of data from the Can-Do-Scale (CDS). The chapter concludes with a summary of the results. SPSS v20.0 was used for all descriptive and inferential analyses.

The purpose of the study was to develop and validate a self-assessment instrument that can be used to obtain a reliable estimate of the foreign language proficiency of native speakers of English. The study also investigated whether a number of factors, (i.e., type of foreign language currently studied, level of education, military branch, military rank, gender, age, immersion experience in the current target language, heritage speaker, prior learning of another foreign language, and qualification at Level 2 or higher) affect students' self-assessment of their ability to speak a foreign language.

Two research questions were explored in this study:

RQ1: What is the relationship between student scores on a self-assessment survey and those on a criterion- referenced speaking proficiency test?

RQ2: What is the impact of various variables on the psychometric properties (validity, reliability and accuracy) of student self-assessment of second language speaking proficiency?

CDS and OPI Pilot Study

The purpose of the pilot study was to test the logistics, validity and reliability of the Can-Do Scale (CDS) and to gather information on the feasibility, prior to the larger study, of avoiding any deficiencies in survey design and improving its quality and efficiency. A small-scale pilot study of the CDS was conducted with students in Category III languages. The students who participated in the pilot study were tested in the context of OPI tester-training workshops. Sixty-one students studying five different languages volunteered to participate in the study: Persian Dari (*n*=22), Persian Farsi (*n*=7), Urdu (*n*=11), Thai (*n*=11) and Tagalog (*n*=10). The length of these category III language courses is 47 weeks, however in the main study, which examined category IV languages, course length was 63 weeks.

DLIFLC offers workshops to train teachers to become OPI testers. One of the certification requirements is that the teacher must perform practice, face-to-face and telephonic OPI tests to show that he/she can conduct valid and reliable tests. The examinees for the practice "mock" OPI tests are selected from different semesters according to the desired level of speaking proficiency that is required for the teachers to experience at the different stages of the OPI certification workshop.

The practice tests were not at the end of the course, and that may have affected the results of the pilot study because the students may not have taken these tests seriously, since they had no effect on final grades or meeting graduation requirements. Thus, the conditions of the pilot study differed from those of the main study in terms of the type of languages examined, point in the course when the OPI tests were conducted, and students normally do not prepare for the mock OPI test.

Subjects who participated in the pilot study completed the CDS before the Oral Proficiency Interview (OPI) was conducted. Each of the oral interviews in the five languages involved in the pilot study was recorded according to DLIFLC's standard administration procedures. Subject to the consent of examinees and workshop participants, the interviews were recorded so that researcher could investigate the relationship between the CDS instrument and the practice OPI tests.

The CDS was scored by the researcher before receiving the OPI scores. The oral interview was scored by the two testers who conducted the test. The rest of the participants in the workshop, acting as third raters, also scored the interview. The final rating of the interview was based on the norming between testers, third raters, and the trainers of the workshop.

In the ***Persian Dari*** pilot study, the ***inter-rater reliability*** of the mock OPI interviews, based on the independent ratings assigned by the two testers and third raters was extremely high. In the 22 interviews in *Persian Dari* there were only 4 interviews in which the two testers assigned different ratings (one 0+/1, one 1/1+, one 1+/2 and one 2/2+). Spearman's rho correlation coefficient was $r = .919$, $p < .05$ and Kendall's tau-b was $r = .6887$, $p < .05$ (see Table 3).

The ***correlation*** between the self-assessment ratings of the CDS and the oral interviews in ***Persian Dari*** suggests a significant relationship. In the 22 OPI interviews in Persian Dari there were only five cases in which there were discrepancies between the CDS and the OPI scores (two 1/1+, one 1/2+, one 1+/2+ and one 2/2+). Spearman's correlation coefficient was $r = .667$, $p < .05$ and Kendall's tau-b was $r = .643$, $p < .05$. Table 2 shows that 17 (77.3%) of the 22 subjects, or77.3% received exactly the same

score on the CDS and OPI. The level of agreement between the CDS and OPI, as measured by Cohen's Kappa, was .657.

In the ***Urdu*** pilot study, the ***inter-rater reliability*** of the mock OPI interviews, based on the independent ratings assigned by the two testers and third raters was significant. In the 11 interviews in *Urdu* there were only two cases in which the two testers assigned different ratings (two 1+/2). The Spearman's correlation coefficient was $r = .638$, $p < .05$ and Kendall's tau-b was $r = .620$, $p < .05$ (see Table 3).

The ***correlation*** between the self-assessment ratings of the CDS and the oral interviews in ***Urdu*** suggests a statistically significant, positive relationship. In the 11 OPI interviews in *Urdu* there was only one case in which there were discrepancies between the CDS and the OPI scores (1+/2). The Spearman's correlation coefficient was $r = .742$, $p < .05$ and Kendall's tau-b was $r = .725$, $p < .05$ Table 2 shows that 10 of the 11 subjects, or 90.9%, received exactly the same score on the CDS and OPI. The level of agreement between the CDS and OPI, as measured by Cohen's Kappa was, .633.

In the ***Persian Farsi*** pilot study, the ***inter-rater reliability*** of the mock OPI interviews, with independent ratings assigned by the testers and third raters was significant. In the 11 interviews in *Persian Farsi* there were only two interviews in which the two testers assigned different ratings (two 1/1+). The Spearman's correlation coefficient was $r = .738$, $p < .05$ and Kendall's tau-b was $r = .723$ $p < .05$ (Table 3).

The ***correlation*** between the self-assessment ratings of the CDS and the oral interviews in ***Persian Farsi*** suggests a statistically significant, positive relationship. In the 8 OPI interviews in ***Persian Farsi*** there were only two cases in which there were discrepancies between the CDS and the OPI scores (1/1+). The Spearman correlation

coefficient was $r = .837$, $p < .05$ and Kendall's tau-b was $r = .811$, $p < .05$. Table 2 shows that, 6 out of the 8 subjects, or 75%, received exactly the same score on the CDS and OPI. The level of agreement between the CDS and OPI, as measured by Cohen's Kappa was, .600.

In the ***Tagalog*** pilot study, the ***inter-rater reliability*** of the mock OPI interviews was perfect. In the 11 face- to face interviews in Tagalog there were no discrepancies between the two testers. Both testers assigned the same rating in all 11 tests. The Spearman's rho correlation coefficient was $r = 1.000$, $p < .05$ and Kendall's tau-b was $r = 1.000$, $p < .05$ (see Table 3).

The ***correlation*** between the self-assessment of the CDS and the oral interviews in ***Tagalog*** suggests a significant and positive relationship. In the 11 OPI interviews in *Tagalog* there were only four cases in which there were discrepancies between the CDS and the OPI scores. The Spearman correlation coefficient was $r = .833$, $p < .05$ and Kendall's tau-b was $r = .778$, $p < .05$ Table 2 shows that, 7 of the 11 subjects, or 63.6%, received exactly the same score on the CDS and OPI. The level of agreement between the CDS and OPI, as measured by Cohen's Kappa was .382.

In the ***Thai pilot*** study, the inter-rater reliability of the mock OPI tests was perfect. In the 10 face-to face interviews in Thai there were no discrepancies between the two testers. Both testers assigned the same rating in all 10 tests. The Spearman's rho correlation coefficient was $r = 1.000$, $p < .05$ and Kendall's tau-b was $r = 1.000$, $p < .05$ (see Table 3).

The ***correlation*** between the self-assessment of the CDS and the oral interviews in ***Thai*** suggests a significant relationship. In the 10 OPI interviews in *Thai* there were

only three cases in which there were discrepancies between the CDS and the OPI scores. The Spearman's correlation coefficient was $r = .830$, $p < .05$ and Kendall's tau-b was $r = .787$, $p < .05$ as table 2 shows that, 7 of the 10 subjects, or 70%, received exactly the same score on the CDS and OPI. The level of agreement between the CDS and OPI, as measured by Cohen's Kappa was, .500. The results of the pilot study showed that there was a significant correlation between CDS and OPI scores, which supports Hypothesis 1.

Tables 2 and 3 summarize the correlation between CDS and OPI and also show the inter-rater reliability of the two testers who rated independently.

Table 2: Correlation between the CDS-and the OPI in the Pilot Study

Language	N	Spearman's	Kendall's tau-b	Cohen's Kappa
Persian Dari	22	.667**	.643**	.657**
Persian Farsi	7	.837**	.811**	.600**
Tagalog	10	.833**	.778**	.382**
Thai	11	.830**	.787**	.500**
Urdu	11	.742**	.725**	.633**

**. Correlation is significant at $p < .05$

Table 3: OPI Inter Rater Reliability in the Pilot Study

Language	N	Spearman's	Kendall's tau-b
Persian Dari	22	.919**	.887**
Persian Farsi	7	.738**	.723**
Tagalog	10	1.000**	1.000**
Thai	11	1.000**	1.000**
Urdu	11	.638**	.620**

**. Correlation is significant at $p < .05$

Population and Demographic Analysis

The main study was conducted at the Defense Language Institute Foreign Language Center (DLIFLC) in Monterey, California. The research population consisted of Basic Course students in three Category IV languages, Arabic, Chinese or Korean. The sample consisted of 350 volunteers from that population.

The sample consisted of 83 females and 267 males. Participants ranked from private to major. The sample consisted of 220 students who studied Arabic, 65 students who studied Chinese, and 65 students who studied Korean. The students who volunteered to participate in the study were from four different military branches: Army, Air Force, Marines, and Navy, and included enlisted service members and commissioned officers. The age of the volunteers ranged from 18 to 40.

The sample of Arabic students consisted of 169 males and 51 females, 203 enlisted soldiers and 17 officers from second lieutenant to major. The sample of Korean students consisted of 49 males and 16 females. The sample of Chinese students consisted

of 49 males and 16 females. There was only one officer in each of the Chinese and Korean samples and 64 enlisted.

Inferential Analysis

Instrumentation

The Can-Do Scale (CDS) was used in the study to examine the relationship between the scores of student self-assessment and the end-of-course Oral Proficiency Interview (OPI). The CDS instrument was adapted from the self-assessment of foreign language speaking proficiency on the ILR website. The CDS went through an intensive validation process in this study, since it did not have any previous testing.

The validation process went through multiple steps: (a) item validation in terms of proficiency level, (b) item validation in terms of being readily comprehensible, (c) test re-tests reliability study, and (d) pilot study. The test re-tests reliability study with different forms of the CDS showed statistical significant results. The pilot study, which was conducted in five Category III languages, indicated that there was a positive relationship between students' self-assessment and OPI ratings.

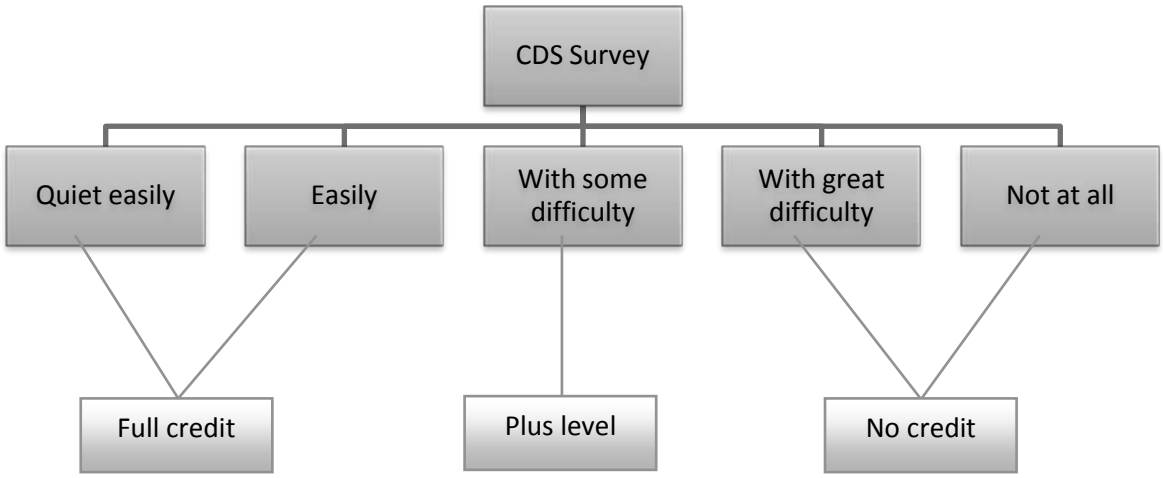
Variables

The CDS instrument collected information on other variables (e.g., type of foreign language currently studied, level of education, military branch, military rank, gender, age, immersion experience in the current target language, heritage speaker, and prior learning of another foreign language) to determine whether they had an effect on students' self-assessment of speaking proficiency.

The instrument included 30 can-do statements. The first 27 items in the survey were stated in the positive, and the Likert type responses were: *"quite easily, easily, with some difficulty, with great difficulty,* and *not at all*".

Although, the CDS is a five-point scale, the rating is based on a three-point scale. An examinee who checks "quite *easily" or "easily"* receives full credit for the base level associated with the can-do item in question. An examinee who checks "*with some difficulty"* receives the plus level. For example, a "with some difficulty" response to a Level 2 can-do statement would be evidence that the examinee might be a 1+.An examinee who checks "*with great difficulty" or "not at all" receives* no credit at all for being at the proficiency level associated with the item (see the diagram below).

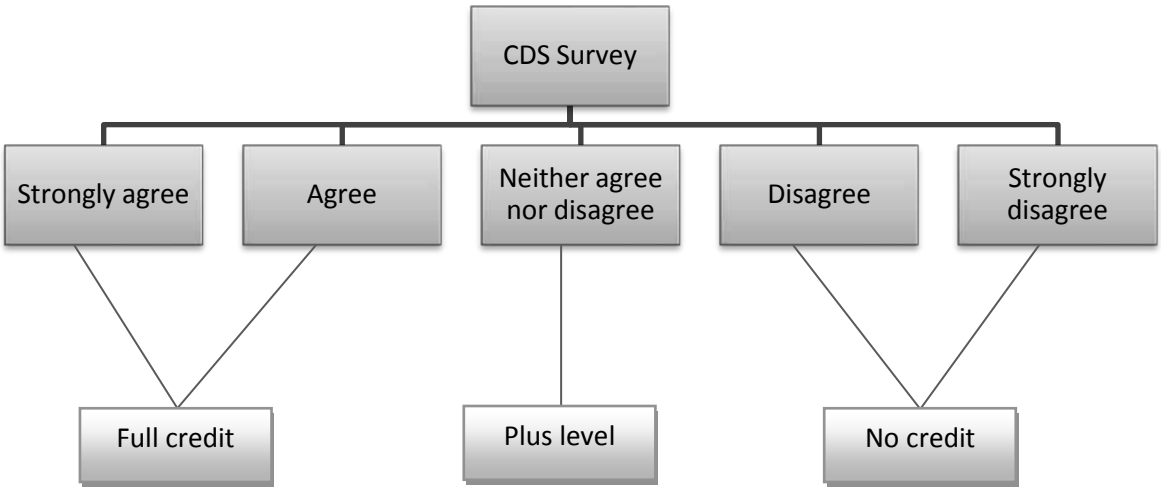
Figure 1: **Survey Scoring Protocols**



The last three items were stated in the positive too, but they have different responses "*strongly agree, agree, neither agree nor disagree, disagree,* and s*trongly*

disagree. An examinee who checks "*strongly agree*," or "agree" receives full credit for the base level. An examinee who checks "n*either agree nor disagree"* receives the plus level. An examinee who checks "*disagree,"* or "s*trongly disagree"* receives no credit (see the diagram below).The instrument was scored according to the scoring rules presented in chapter 3(see Table 1). A complete copy of the instrument is provided in Appendix E.

Figure 2: **Survey Scoring Protocols**



Responses to the CDS were collected and coded numerically to assure accuracy in data entry while assuring participant confidentiality. Participant names or other personally identifying information were not retained for data analysis and reporting purposes. Confidentiality was maintained in compliance with applicable federal regulations and statutes. The students were informed that the limits of confidentiality might be broken in the case of a court subpoena, or other lawful means.

CDS Test-Retest Reliability

Test-retest reliability refers to the test's stability among different administrations. The CDS was administered with an alternate form on two separate occasions to a group of Arabic students in three different semesters. Two forms (A & B) of the CDS were developed from the same set of items by presenting the items in a different order. The order of the items was scrambled within each proficiency level. Form A was administered to a group of students and in approximately seven to ten days form B was administered to the same group. The reliability test population consisted of 40 students in semester I, 16 in semester II, and 24 in semester III.

The results showed that the test-retest reliability for the semester I students was statistically significant ($r = .638$, $p < .05$) (see Table 4). In semester II the results showed that the test was statistically significant. ($r = .902$, $p < .05$) (see Table 5). In semester III the results also showed that the test was statistically significant. ($r = .803$, $p < .05$) (see Table 6). The correlation was lower for the first semester than semester II and III, because at this stage of the course students may not understand fully the ILR descriptors and what are the linguistic factors (e.g., vocabulary, grammar, delivery) requirement to achieve a specific level.

The CDS was shown to be reliable since the scores that each student received on the first administration were consistent their scores on the second. As noted above, the correlation between the scores from the two administrations was statistically significant in each of the semesters. Students in semester II had the highest reliability ($r = .902$.)

These results indicated that the CDS was a reliable instrument that could be used in the main study.

Table 4: Test-Retest correlation Semester I

Spearman's rho	Tester 1 (Form A)	Tester 2(Form B)
N	40	40
Correlation Coefficient	1.000	.638**
Sig. (2-tailed)	.	.000

**. Correlation is significant at p < .05

Table 5: Test-Retest correlation Semester II

Spearman's rho	Tester 1 (Form A)	Tester 2(Form B)
N	16	16
Correlation Coefficient	1.000	.902**
Sig. (2-tailed)	.	.000

**. Correlation is significant at p < .05

Table 6: Test-Retest correlation Semester III

Spearman's rho	Tester 1 (Form A)	Tester 2(Form B)
N	24	24
Correlation Coefficient	1.000	.803**
Sig. (2-tailed)	.	.000

**. Correlation is significant at p < .05

Inter-rater reliability of CDS Raters

The inter-rater reliability of the original two raters of the CDSs was extremely high in the entire sample (n=350) of Arabic, Chinese and Korean students. The third rating was conducted independently by a single master tester. The Spearman's rho correlation coefficient between the original, independent ratings of the two testers who scored the CDSs was ($r = .943$, $p < .05$). Kendall's tau_b was ($r = .926$, $p < .05$), as presented in Table 7**.**

Table 7: Inter-rater Reliability of CDS Raters

			First Rater Survey Score	Second Rater Survey Score
Kendall's tau_b	First Rater Survey Score	Correlation Coefficient	1.000	.926**
		Sig. (2-tailed)	.	.000
		N	350	350
	Second Rater Survey Score	Correlation Coefficient	.926**	1.000
		Sig. (2-tailed)	.000	.
		N	350	350
Spearman's rho	First Rater Survey Score	Correlation Coefficient	1.000	.943**
		Sig. (2-tailed)	.	.000
		N	350	350
	Second Rater Survey Score	Correlation Coefficient	.943**	1.000
		Sig. (2-tailed)	.000	.
		N	350	350

**. Correlation is significant at $p < .05$

Inter-rater reliability of OPI Raters

The Spearman's correlation coefficient between the original, independent ratings of the two testers who scored the OPIs was $r = .975$, $p < .05$. The inter-rater reliability of the original two raters of the OPIs was exceptionally high in the entire sample (n=350) of Arabic, Chinese and Korean students. The third ratings were conducted independently by single master testers.

Kendall's tau_b was ($r = .973$, $p < .05$), as presented in Table 8. In the majority of the OPIs (97.43%) the two testers assigned exactly the same rating. When the two ratings for a given student differed, they were typically within a plus level. 164 (47.0%) of the 350 tests were third rated for quality control. In 145 of 164 tests (88.4%) the third raters agreed with the initial raters.

Table 8: Inter-rater reliability of OPI Raters

			First Rater OPI Score	Second Rater OPI Score
Kendall's tau_b	First Rater OPI Score	Correlation Coefficient	1.000	.973**
		Sig. (2-tailed)	.	.000
		N	350	350
	Second Rater OPI Score	Correlation Coefficient	.973**	1.000
		Sig. (2-tailed)	.000	.
		N	350	350
Spearman's rho	First Rater OPI Score	Correlation Coefficient	1.000	.975**
		Sig. (2-tailed)	.	.000
		N	350	350
	Second Rater OPI Score	Correlation Coefficient	.975**	1.000
		Sig. (2-tailed)	.000	.
		N	350	350

**. Correlation is significant at $p < .05$

Research Question 1

Hypothesis Test

To assess the relationship between students' self-assessment of their ability in speaking foreign languages and their official Oral proficiency Interview (OPI) scores, the following hypothesis was tested and the following research question was asked:

H1: There is a relationship between student scores on a self-assessment survey and those on a criterion-referenced speaking proficiency test.

$H1_0$: There is no relationship between student scores on a self-assessment survey and those on a criterion-referenced speaking proficiency test.

RQ1: What is the relationship between student scores on a self-assessment survey and those on a criterion-referenced speaking proficiency test?

Viewing the ratings as ordinal and nominal, rather than interval, the researcher used SPSS to compute a Spearman's rho and a Kendall's tau_b correlation coefficient. To address this question, correlations were computed using the nonparametric measures of Spearman's rho and Kendall's tau_b. Spearman's rho, like Kendall's tau_b rank correlation (Crichton, 1999), is carried out on the ranks of the data. That is, for each variable separately the values are put in order and numbered, 1 for the lowest value, 2 for the next value and so on.

An examination of Table 9 indicates a significant relationship between the self-assessment measured by the CDS instrument and the final OPI scores. The data in Table 9 present the outcome of the entire sample in the three languages tested in this study: Arabic, Chinese Mandarin, and Korean (n=350). Kendall's tau-b was $r = .260$, $p < .05$ and Spearman's rho was $r = .272$ $p < .05$. The result is significant but is not strong. A separate data analysis for each language will follow in this chapter.

Table 9: Correlations of CDS and OPI for the entire Sample

			Survey Final Rating	OPI Final Rating
Kendall's tau_b	Survey Final Rating	Correlation Coefficient	1.000	.260**
		Sig. (2-tailed)	.	.000
		N	350	350
	OPI Final Rating	Correlation Coefficient	.260**	1.000
		Sig. (2-tailed)	.000	.
		N	350	350
Spearman's rho	Survey Final Rating	Correlation Coefficient	1.000	.272**
		Sig. (2-tailed)	.	.000
		N	350	350
	OPI Final Rating	Correlation Coefficient	.272**	1.000
		Sig. (2-tailed)	.000	.
		N	350	350

**. Correlation is significant at the 0.05 level (2-tailed).

Table 10 presents the relationship between CDS and OPI scores for the Arabic (AD) students (*n*=220). The table reports a low but significant relationship between students' self-assessment and their final OPI scores. Kendall's tau_b correlation coefficient was (r = .276 p < .05), and Spearman's rho was (r = .289 p < .05).

Table 10: Correlations of the CDS and OPI for Arabic (AD) Students

			OPI Final Rating	Survey Final Rating
Kendall's tau_b	OPI Final Rating	Correlation Coefficient	1.000	.276**
		Sig. (2-tailed)	.	.000
		N	220	220
	Survey Final Rating	Correlation Coefficient	.276**	1.000
		Sig. (2-tailed)	.000	.
		N	220	220
Spearman's rho	OPI Final Rating	Correlation Coefficient	1.000	.289**
		Sig. (2-tailed)	.	.000
		N	220	220
	Survey Final Rating	Correlation Coefficient	.289**	1.000
		Sig. (2-tailed)	.000	.
		N	220	220

**. Correlation is significant at the 0.05 level (2-tailed).

Table 11 presents the relationship between CDS and OPI scores for the Chinese Mandarin (CM) students (*n*=65). The table reports no statistically significant relationship between students' self-assessment and their final OPI scores. Kendall's tau_b correlation coefficient was (r = .195 p < .05), and Spearman's rho was (r = .208 p < .05).

Table 11: Correlations of the CDS and OPI for Chinese Mandarin (CM) Students

			OPI Final Rating	Survey Final Rating
Kendall's tau_b	OPI Final Rating	Correlation Coefficient	1.000	.195
		Sig. (2-tailed)	.	.097
		N	65	65
	Survey Final Rating	Correlation Coefficient	.195	1.000
		Sig. (2-tailed)	.097	.
		N	65	65
Spearman's rho	OPI Final Rating	Correlation Coefficient	1.000	.208
		Sig. (2-tailed)	.	.097
		N	65	65
	Survey Final Rating	Correlation Coefficient	.208	1.000
		Sig. (2-tailed)	.097	.
		N	65	65

**. Correlation is significant at the 0.05 level (2-tailed).

Table 12 presents the relationship between CDS and OPI scores for the Korean (KP) students (*n*=65). The table reports a positive relationship between students' self-assessment and their final OPI scores. Kendall's tau_b correlation coefficient was (r = .240 $p < .05$), and Spearman's rho was (r = .249 $p < .05$). Both correlations were significant but weak.

Table 12: Correlations of the CDS and OPI for Korean (KP) Students

			OPI Final Rating	Survey Final Rating
Kendall's tau_b	OPI Final Rating	Correlation Coefficient	1.000	.240*
		Sig. (2-tailed)	.	.044
		N	65	65
	Survey Final Rating	Correlation Coefficient	.240*	1.000
		Sig. (2-tailed)	.044	.
		N	65	65
Spearman's rho	OPI Final Rating	Correlation Coefficient	1.000	.249*
		Sig. (2-tailed)	.	.045
		N	65	65
	Survey Final Rating	Correlation Coefficient	.249*	1.000
		Sig. (2-tailed)	.045	.
		N	65	65

*. Correlation is significant at the 0.05 level (2-tailed).

Table 13 presents a crosstabulation analysis of CDS-survey and OPI scores for the entire sample in the three languages--Arabic, Chinese Mandarin and Korean. Table 13 is intended to answer the question of the extent to which CDS scores agreed perfectly with the criterion OPI scores.

The researcher decided to exclude Level 2+ from this analysis because there were so few students with OPI scores of 2+ that percent agreement statistics would not be meaningful. The ILR scale consists of an 11-point scale from level 0+ to level 5; however, for this study the CDS survey scores reached only level 3. The students who

participated were in the Basic-Course Program and scored no higher than Level 2+ on the official end-of-course OPI.

The level of exact agreement for the entire sample was 58.3%: 204 of the 350 students received exactly the same score on the CDS and OPI. The table shows the highest level of agreement among scores near the lower end of the ILR scale. That is, the highest level of percentage agreement between CDS and OPI scores occurred at Level 1+. The level of exact agreement at Level 1+ was 71.6%: 154 of the 215 students received a 1+ on both measure. The second highest level of agreement was at Level 2; where out of the 131 students who earned a Level 2 on the OPI, 48 (36.6 %) received a 2 on the CDS.

Table 13: OPI and CDS Crosstabulation for the Entire Sample (AD-CM-KP)

			Survey Final Rating						
			0+	1	1+	2	2+	3	Total
OPI Final Rating	1+	Count	1	3	154	34	23	0	215
		% within OPI Final Rating	.5%	1.4%	71.6%	15.8%	10.7%	.0%	100.0%
		% of Total	.3%	.9%	44.0%	9.7%	6.6%	.0%	61.4%
	2	Count	0	0	63	48	17	3	131
		% within OPI Final Rating	.0%	.0%	48.1%	36.6%	13.0%	2.3%	100.0%
		% of Total	.0%	.0%	18.0%	13.7%	4.9%	.9%	37.4%
	2+	Count	0	0	0	1	2	1	4
		% within OPI Final Rating	.0%	.0%	.0%	25.0%	50.0%	25.0%	100.0%
		% of Total	.0%	.0%	.0%	.3%	.6%	.3%	1.1%
	Total	Count	1	3	217	83	42	4	350
		% within OPI Final Rating	.3%	.9%	62.0%	23.7%	12.0%	1.1%	100.0%
		% of Total	.3%	.9%	62.0%	23.7%	12.0%	1.1%	100.0%

Table 13 (above) shows an overview of the entire sample and to facilitate comprehension, the same information is presented in a bar-chart form in Figure 3.

Figure 3: Bar Chart Crosstabulation CDS/OPI for the Entire Sample (N=350) (AD-CM-KP)

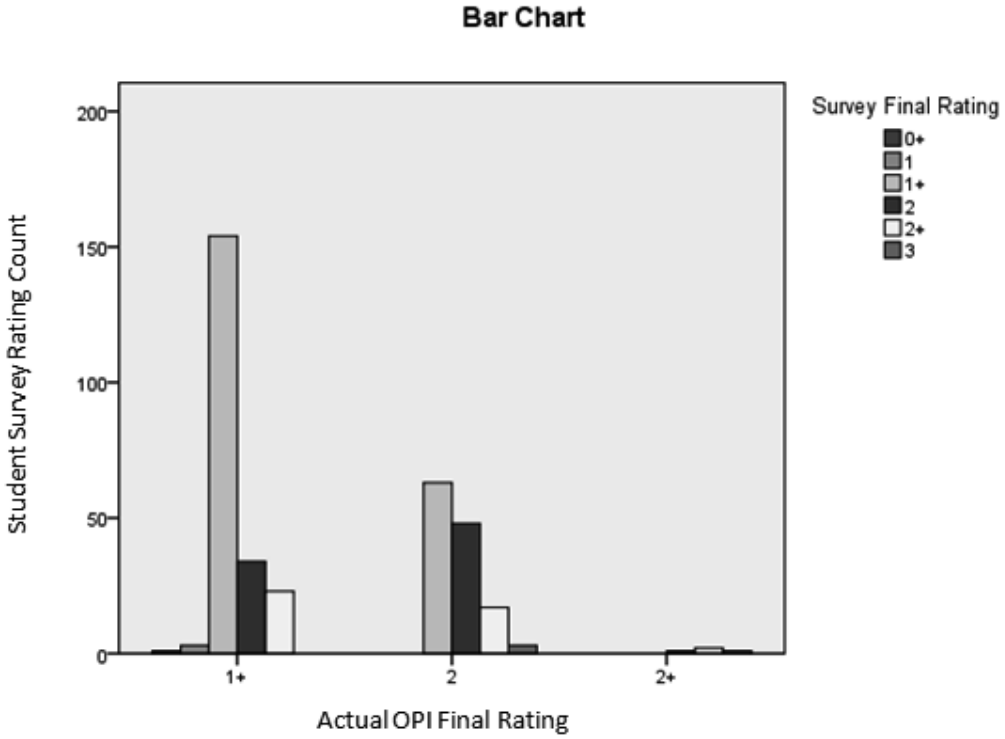


Table 14 is intended to answer the question of the extent to which CDS scores agreed perfectly with the criterion OPI scores for Arabic students. Table 14 shows that students have the highest level of agreement between CDS and OPI scores at level 1+. Level 2+ was excluded from this analysis because there were so few students with OPI scores of 2+ that percent agreement statistics would not be meaningful. As shown in the Table 14, 151 received Level 1+ on the final OPI score; 105 of the 151 students, or (69.5 %) rated their language proficiency at 1+. The second highest level of agreement was at Level 2, where 68 received Level 2 on the final OPI measure. Thirty of the 68 students, or (44.1 %) rated their language proficiency at 2.

Table 14: OPI and CDS Crosstabulation for Arabic (AD) Students

			Survey Final Rating						
			0+	1	1+	2	2+	3	Total
OPI Final Rating	1+	Count	1	2	105	25	18	0	151
		% within OPI Final Rating	.7%	1.3%	69.5%	16.6%	11.9%	.0%	100.0%
		% of Total	.5%	.9%	47.7%	11.4%	8.2%	.0%	68.6%
	2	Count	0	0	27	30	10	1	68
		% within OPI Final Rating	.0%	.0%	39.7%	44.1%	14.7%	1.5%	100.0%
		% of Total	.0%	.0%	12.3%	13.6%	4.5%	.5%	30.9%
	2+	Count	0	0	0	0	1	0	1
		% within OPI Final Rating	.0%	.0%	.0%	.0%	100.0%	.0%	100.0%
		% of Total	.0%	.0%	.0%	.0%	.5%	.0%	.5%
	Total	Count	1	2	132	55	29	1	220
		% within OPI Final Rating	.5%	.9%	60.0%	25.0%	13.2%	.5%	100.0%
		% of Total	.5%	.9%	60.0%	25.0%	13.2%	.5%	100.0%

Figure 4: Bar Chart Crosstabulation CDS/OPI for Arabic Students

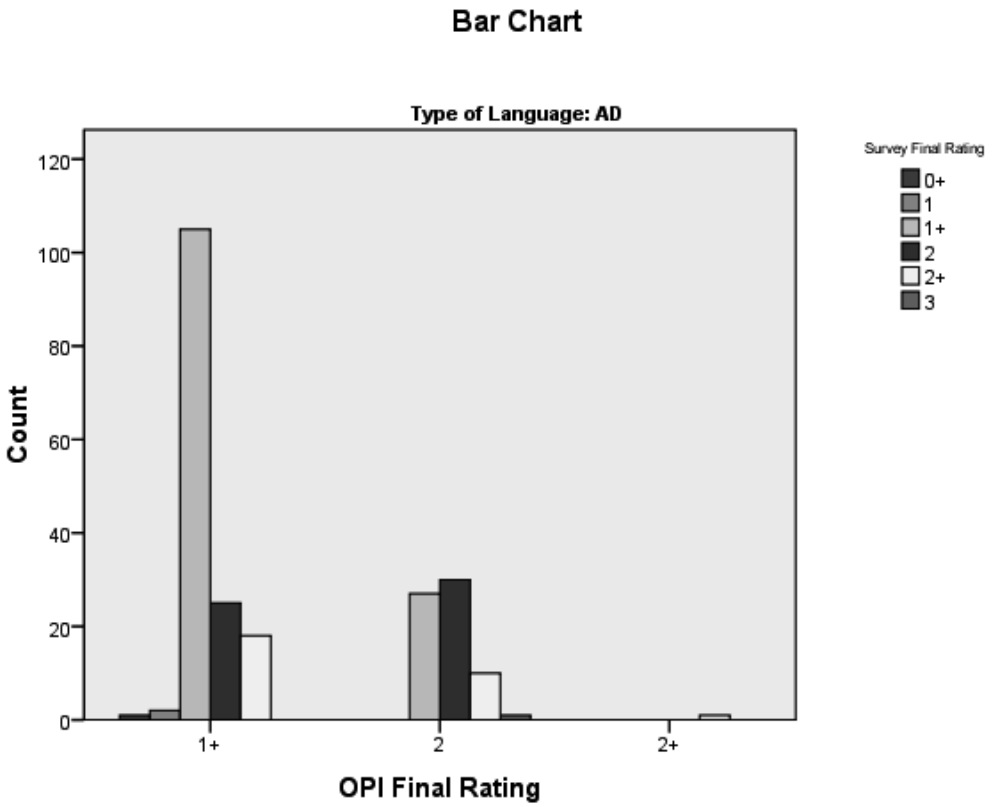


Table 15 is intended to answer the question of the extent to which CDS scores agreed perfectly with the criterion OPI scores for Chinese Mandarin (CM) students. Table 15 shows that students have the highest level of agreement between CDS and OPI scores at level 1+. As can be seen in the table, 24 students received Level 1+ on the final OPI, and 17 of the 24 students (70.8 %) received 1+ on the CDS. The second highest level of agreement was at Level 2, where 40 students received Level 2 on the final OPI, and fourteen of the 40 students (35.0 %) received 2 on the CDS. Level 2+ was excluded from this analysis because there were so few students with OPI scores of 2+ that percent agreement statistics would not be meaningful.

Table 15: OPI and CDS Crosstabulation for Chinese Mandarin (CM) Students

OPI Final Rating		Survey Final Rating				
		1+	2	2+	3	Total
1+	Count	17	3	4	0	24
	% within OPI Final Rating	70.8%	12.5%	16.7%	.0%	100.0%
	% of Total	26.2%	4.6%	6.2%	.0%	36.9%
2	Count	19	14	5	2	40
	% within OPI Final	47.5%	35.0%	12.5%	5.0%	100.0%
	Rating% of Total	29.2%	21.6%	7.7%	3.1%	61.5%
2+	Count	0	1	.0%	0	1
	% within OPI Final	.0%	100.0	.0%	.0%	100%
	Rating% of Total	.0%	1.5%		.0%	1.5%
Total	Count	36	18	9	2	65
	% within OPI Final	55.4%	27.7%	13.8%	3.1%	100.0%
	Rating% of Total	55.4%	27.7%	13.8%	3.1%	100.0%

An obvious question about the CDS is the extent to which it yields under- or over-estimates. Table 15 indicates that Chinese Mandarin students are underestimating their proficiency as measured by the OPI. As can be seen in the table, 36 students received 1+ on the CDS, while 19 of the 36 (52.8%) underestimated their proficiency and received level 2 on the final OPI.

The percentage of perfect agreement for Levels 1+ and 2 in the Arabic language was comparable to the percentage among these levels for the entire sample (see Table 16) For Level 1+ there was a comparable percentage of perfect agreement between Arabic and Chinese: 69.5% for Arabic compared with 70.8% for Chinese. On the other hand,

there was a striking difference in favor of Arabic at level 2. Level 2 in Arabic was (44.1%) compared with (35.0%) for Chinese Mandarin (See Tables 14, 15 and Figure 5).

For Level 1+ there was a noticeable difference in the percentage of perfect agreement between Chinese and Korean: (80.0%) for Korean compared with (70.8%) for Chinese. On the other hand, there was a significant difference in favor of Chinese at level 2, (35.0%) for Chinese compared with (17.4%) for Korean (see Table 17 and Figure 6).

Figure 5: **Bar Chart Crosstabulation CDS/OPI for Chinese Mandarin Students**

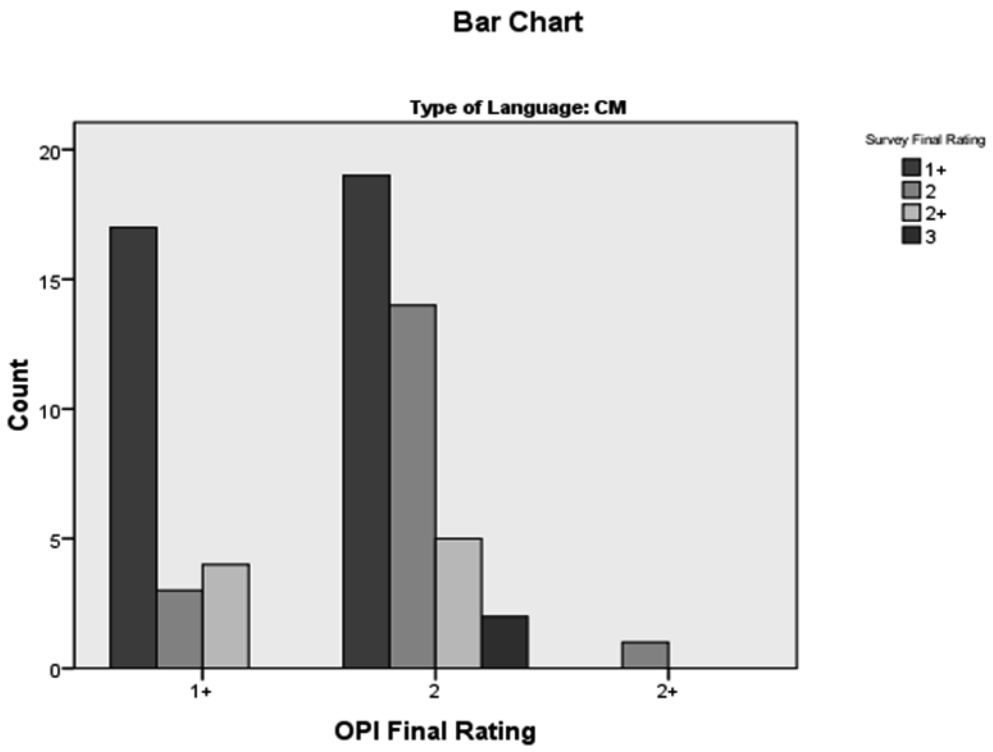


Table 16: Perfect Agreement between CDS and OPI: Comparison of Arabic, Chinese and Korean with Entire Sample

Language	Level 1+	Level 2
Arabic	69.5%	44.1%
Chinese Mandarin	70.8%	35.0%
Korean	80.0%	17.4%
Entire Sample	71.6%	36.6%

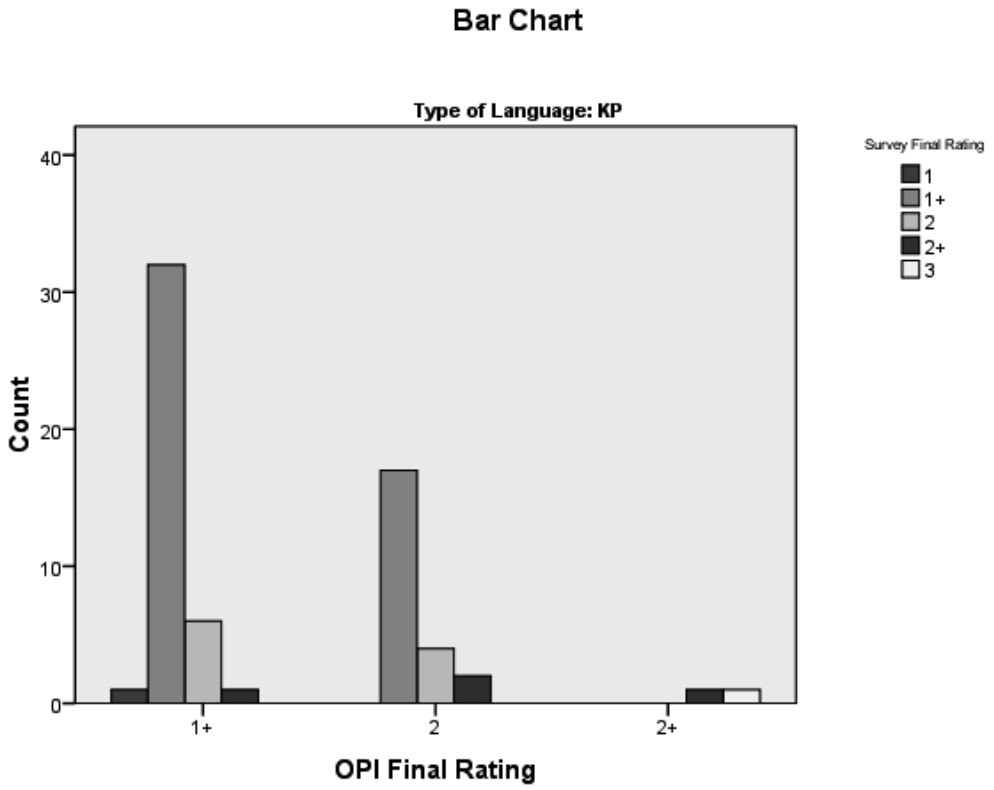
Table 17 is intended to answer the question of the extent to which CDS scores agreed perfectly with the criterion OPI scores for Korean (KP) students. Table 17 shows that students have the highest level of agreement between CDS and OPI scores at level 1+. As can be seen in the table, 40 students received Level 1+ on the final OPI. Thirty two out of the 40 students (80.0 %) received 1+ on the CDS.

The second highest level of agreement was at Level 2, where 23 students received Level 2 on the final OPI. Four of the 23 students (17.4 %) received 2 on the CDS. Level 2+ was excluded from this analysis because there were so few students with OPI scores of 2+ that percent agreement statistics would not be meaningful.

Table 17 indicates that17 of the 49 students (34.7%) who received a 1+ on the CDS underestimated their proficiency as measured by the OPI (Level 2). Six of the 10 students (60.0%) who received a 2 on the CDS overestimated their proficiency as measured by the OPI (Level 1+).

Table 17: OPI and CDS Crosstabulation for Korean (KP) Students

OPI Final Rating		Survey Final Rating					
		1	1+	2	2+	3	Total
1+	Count	1	32	6	1	0	40
	% within OPI Final Rating	2.5%	80.0%	15.0%	2.5%	.0%	100.0%
	% within Survey Final Rating	100.0%	65.3%	60.0%	25.0%	.0%	61.5%
2	Count	0	17	4	2	0	23
	% within OPI Final Rating	.0%	73.9%	17.4%	8.7%	.0%	100.0%
	% within Survey Final Rating	.0%	34.7%	40.0%	50.0%	.0%	35.4%
2+	Count	0	0	0	1	1	2
	% within OPI Final Rating	.0%	.0%	.0%	50.0%	50.0%	100.0%
	% within Survey Final Rating	.0%	.0%	.0%	25.0%	100.0%	3.1%
Total	Count	1	49	10	4	1	65
	% within OPI Final Rating	1.5%	75.4%	15.4%	6.2%	1.5%	100.0%
	% within Survey Final Rating	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%



CDS/OPI Perfect Agreement

The correlation between CDS and OPI scores was statistically significant but low in Arabic and Korean languages. However, the data analysis showed no statistically significant correlation between CDS and OPI scores in Chinese, which addresses the first part of research question 1.

RQ1: What is the relationship between student scores on a self-assessment survey and those on a criterion- referenced speaking proficiency test?

The second part of the research question 1 investigates how often examinees received the same rating on both the CDS and OPI (Perfect Agreement). As shown in Table 18, there is a moderate level of agreement in the three languages. In Arabic, 136 of the 220 students, or (61.8%) received exactly the same score on the CDS and OPI.

In Korean, 37 of the 65 students, or (56.9%) received exactly the same score on the CDS and OPI. In Chinese Mandarin, 31of the 65 students, or (47.7%) received exactly the same score on the CDS and OPI. As it is not possible to test percent-agreement for statistical significance, it is important to determine what minimum level of agreement is necessary to infer that the instrument's reliability has been established.

Table 18: Perfect OPI/CSD Agreement for Arabic, Chinese and Korean

			Type of Language			
			AD	CM	KP	Total
Perfect OPI/CSD Agreement	NO	Count	84	34	28	146
		% within Perfect OPI/CSD Agreement	57.5%	23.3%	19.2%	100.0%
		% within Type of Language	38.2%	52.3%	43.1%	41.7%
		% of Total	24.0%	9.7%	8.0%	41.7%
	YES	Count	136	31	37	204
		% within Perfect OPI/CSD Agreement	66.7%	15.2%	18.1%	100.0%
		% within Type of Language	61.8%	47.7%	56.9%	58.3%
		% of Total	38.9%	8.9%	10.6%	58.3%
	Total	Count	220	65	65	350
		% within Perfect OPI/CSD Agreement	62.9%	18.6%	18.6%	100.0%
		% within Type of Language	100.0%	100.0 %	100.0 %	100.0%
		% of Total	62.9%	18.6%	18.6%	100.0%

The data analysis for research question 1 indicated there was a significance correlation between CDS and OPI (r = .272 $p < .05$) (see Table 9). The percentage of

perfect agreement between CDS/OPI for the entire sample was (58.3%), and for the majority of students who had ratings discrepancies the two ratings differed by only a plus level (34.0%). The null hypothesis stated there was no relationship between student scores on a self-assessment survey and those on a criterion- referenced speaking proficiency test. Thus, the null hypothesis was rejected.

CDS/OPI within Range

The students who received different scores on the CDS and OPI had no perfect agreement between both measures when the ILR scale is employed. These students had to be categorized based on how far the CDS and OPI scores differed. If the student scores above or below a plus level in relation to the OPI score, he/she is considered to be within range of the target base level.

For example, a student who measures his/her proficiency on the CDS survey at Level 2 and receives 1+ on the final OPI considered to be within range by a plus level in relation to the base level 2. In this case, the student overestimates his/her proficiency by a plus level. On the other hand, a student who measures his/her proficiency on the CDS survey at Level 1+ and receives level 2 on the final OPI considered to be within range by a plus level in relation to the base level 2. In this case, the student underestimates his/her proficiency by a half level. As shown in Table 19, in Arabic, 64 out of the 220 students, or (29.1%), the CDS and OPI scores were differed by a plus level. 136 out of 220 students, or (61.8%) received exact agreements on both measures (see Table 19), and 20 students, or (9.09%) received different score that higher or lower by a full level.

In Chinese Mandarin, 28 of the 65 students, or (43.1%), had CDS and OPI scores that differed by a plus level. In Korean, 27 of the 65 students, or (41.5%), had

CDS and OPI scores that differed by a plus level. The total number of students in the three languages who had above or below a plus-level discrepancy between their CDS and OPI scores was 119 of the 350 students, or (34%) as shown in Table19.

Table 19: Within Range OPI/CSD Agreement (+/-) Crosstabulation for AD-CM KP

			Type of Language			
			AD	CM	KP	Total
In Range OPI/CSD Agreement (+/-)	NO	Count	156	37	38	231
		% within In Range OPI/CSD Agreement (+/-)	67.5%	16.0%	16.5%	100.0%
		% within Type of Language	70.9%	56.9%	58.5%	66.0%
	YES	Count	64	28	27	119
		% within In Range OPI/CSD Agreement (+/-)	53.8%	23.5%	22.7%	100.0%
		% within Type of Language	29.1%	43.1%	41.5%	34.0%
	Total	Count	220	65	65	350
		% within In Range OPI/CSD Agreement (+/-)	62.9%	18.6%	18.6%	100.0%
		% within Type of Language	100.0%	100.0%	100.0%	100.0%

CDS/OPI outside the Range

If a student's CDS and OPI scores differ by a whole level or more, the student is considered to be "outside the range." For example, a student who measures his/her proficiency on the CDS survey at Level 2+ and receives 1+ on the final OPI is considered as being outside the range by a whole level. In this case, the student

overestimates his/her proficiency by a whole level. On the other hand, a student who measures his/her proficiency on the CDS survey at Level 1+ and receives level 2+ on the final OPI is considered also to be outside the range by a whole level. In this case, the student underestimates his/her proficiency by whole level.

Table 20: OPI/CSD outside the Range (AD-CM-KP)

CSD/OPI Agreement Categories				OPI Final Rating			
				1+	2	2+	Total
Outside Range	Survey Final Rating	0+	Count	1	0		1
			% within OPI Final Rating	4.2%	.0%		3.7%
		2+	Count	23	0		23
			% within OPI Final Rating	95.8%	.0%		85.2%
		3	Count	0	3		3
			% within OPI Final Rating	.0%	100.0%		11.1%
		Total	Count	24	3		27
			% within Survey Final Rating	88.9%	11.1%		100.0%

In Arabic, 19 of the 220 students (8.06%) overestimated their foreign language proficiency by a full level or more. For example, 18 (62.1%) of the 29 students who received a 2+ on the CDS received a 1+ on their final OPI and the one who received a 3 on the CDS and a 2 on the OPI. (see Table 14). In Chinese, 6 of the 65 students, or (9.2%), overestimated their proficiency by a whole level or more. In Korean, only 1 out

of the 65 students, or (1.5%) was overestimated their proficiency by a whole level or more. A total of 27 of the 350 students, or (7.7%), overestimated by a whole level according to the CDS ratings.

Research Question 2

Hypothesis Test

The objective of this question was identification of variables which have a significant impact on how well students can self-assess their speaking proficiency. That is, the study investigated variables which may contribute significantly to perfect agreement between CDS and OPI. The methodology involved developing a model showing the extent to which certain variables affect the ability of students of Arabic, Chinese and Korean languages at the DLIFLC to accurately estimate their proficiency in speaking. The estimate came from a self-assessment instrument (CDS), and was compared with a final OPI score which is the DLIFLC's criterion measure of speaking ability. The following are the second hypothesis and research question:

H2: There is an impact of various variables on the psychometric properties (validity, reliability and accuracy) of student self-assessments of second language speaking proficiency.

$H2_0$**:** There is no impact of various variables on the psychometric properties (validity, reliability and accuracy) of student self-assessments of second language speaking proficiency.

RQ2: What is the impact of various variables on the psychometric properties (validity, reliability and accuracy) of student self- assessments of second language speaking proficiency?

The Models

The goal of this analysis was to design a model which would indicate the combination of variables that was likely to predict perfect agreement between CDS/OPI scores, yet was parsimonious. A stepwise and a simultaneous logistic regression were employed in this study. The dependent variable in the logistic regression, in this case OPI and CDS agreement, is dichotomous. The dependent variable was assigned the value 1 for success (perfect agreement), and the value 0 for failure (lack of agreement). Logistic regression does not make an assumption of normal distribution of the independent variables. The relationship between the predictor and response variables is not a linear function in logistic regression; instead, the logistic regression function is used, which is the log transformation of θ: (Agresti, 1996).

$$\theta = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}$$

Where α = the constant of the equation and, β = the coefficient of the predictor variables. An alternative form of the logistic regression equation is:

$$\text{logit}\left[\theta(\mathbf{x})\right] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

Two models were used to analyze and address question 2. The first model was a backward stepwise regression, that included all subjects (*n* =350) and all variables (which were a mixture of categorical and ordinal variables). This model was selected because it is used in the exploratory phase of research (Menard, 1995), and a few variables in this study were not supported by the literature. Backward stepwise regression was used in the

study to eliminate the variables that predicted minimum effect while keeping variables that had a significant effect on the level of agreement between CDS and OPI scores.

The likelihood–ratio test was employed to test the predictive ability of each independent variable. The likelihood-ratio test used the ratio of the maximized value of the likelihood function for the full model (L_1) over the maximized value of the likelihood function for the simpler model (L_0). The likelihood-ratio test statistic equals:

$$-2\log\left(\frac{L_0}{L_1}\right) = -2[\log(L_0) - \log(L_1)] = -2(L_0 - L_1)$$

This log transformation of the likelihood functions yielded a chi-square statistic. This was the test statistic recommended by DLIFLC research experts when building a model through backward stepwise elimination. Agresti (1996) states that the likelihood-ratio test is more reliable for small sample sizes than the Wald test. The model was tested after the elimination of each variable to ensure that the model still adequately fit the data. The elimination process continued until no more variables could be eliminated. Once the model was fitted, likelihood ratio tests were used to test the significance of the model as a whole and to eliminate variables which did not have predictive ability (Agresti, 1996).

The second model was a simultaneous regression which included only the subjects in Chinese and Korean (*n*=130) and those variables that emerged from the backward stepwise. The simultaneous model was used to analyze the effects of selected predictor variables. The reasons for selecting this model were: 1) an additional variable was added to the model (*attended in-country immersion*); 2) Arabic students did not participate in an in-country immersion and all were excluded from the model; 3) a

number of variables were identified in the stepwise model that had a high level of predictive accuracy; and 4) the researcher decided to include gender in the model, even though it had been eliminated from the first model.

Backward Stepwise Regression of the Entire Sample

Restatement of research question 1:

RQ 1: What is the relationship between student scores on a self-assessment survey and those on a criterion- referenced speaking proficiency test?

Research question 1 finding: there was a significant relationship between student self-assessment as measured by the CDS survey and the OPI. The stepwise model included the entire sample (n =350) of the three languages, Arabic, Chinese Mandarin, and Korean.

The stepwise model consisted of all 10 independent variables: (1) Immersion in Target Country, (2) Gender, (3) Age, (4) Education Level, (5) Military Rank, (6) Military Branch, (7) Prior Experience in Studying Foreign language(s), (8) Type of Foreign Language, (9) Heritage Speaker, and (10) Students Who Qualified at ILR Level 2 or Higher. A backward elimination procedure was used to examine which predictors or combination of predictors seemed to have a predictive impact on the perfect agreement between CDS and OPI.

Table 21: Backward Stepwise Model Level of Accuracy

Observed			Predicted		
			Perfect OPI/CDS Agreement		
			No	Yes	Percentage Correct
Step 0	Perfect OPI/CDS Agreement	NO	0	146	.0
		Yes	0	204	100.0
		Overall Percentage			58.3

Then, subsets of predictor variables which showed the least significance were removed and the model was run again. This iterative procedure was continued until a satisfactory significant model was obtained with high predictive accuracy. Reaching a satisfactory model is a matter of researcher judgment based on the accuracy (58.3%) of the model to predict the perfect agreement. Tables 21 and 24 present a comparison between the current full model (all variables) and the following reduced model for Chinese and Korean students (selected variables) to examine whether there was a significant difference between them (Agresti, 1996).

Table 22: Categorical Variables in the Full Stepwise Model

Variable Description		Frequency	Parameter coding			
			(1)	(2)	(3)	(4)
Duration of Previous Language Training	No Training	60	1.000	.000	.000	.000
	6 Months or Less	18	.000	1.000	.000	.000
	6 Months to 12 Months	37	.000	.000	1.000	.000
	12 to 24 Months	75	.000	.000	.000	1.000
	More than 24 Months	160	.000	.000	.000	.000
Military Branch	ARMY	138	1.000	.000	.000	
	AIR FORCE	105	.000	1.000	.000	
	MARINES	56	.000	.000	1.000	
	NAVY	51	.000	.000	.000	
Age Category	18-20	70	1.000	.000		
	21-25	180	.000	1.000		
	26 and Over	100	.000	.000		
Type of Language	AD	220	.000	.000		
	CM	65	1.000	.000		
	KP	65	.000	1.000		
Education Level	GED/High School	225	1.000	.000		
	Associate Degree	41	.000	1.000		
	College Grad	84	.000	.000		
Heritage Speaker	NO	316	1.000			
	YES	34	.000			
Indicated Multiple Language Training	NO	222	1.000			
	YES	128	.000			
History of Language Training	NO	60	1.000			
	YES	290	.000			
Gender	Female	83	1.000			
	Male	267	.000			
Rank	Officer	20	1.000			
	Enlisted	330	.000			
OPI Qualification at ILR=2	Not Qualified	215	1.000			
	Qualified	135	.000			

Table 22 show the demographic categorical variables: duration of previous language training, type of language training, multiple language training, military branch, age, education level, heritage speaker, gender, rank, and OPI Qualification at ILR level 2. Additionally, Table 22 presents 31 subsets of predictor variables.

Table 23 shows the results of the backward stepwise regression. The model indicated that the majority of the variables that had no ability to predict perfect agreement between CDS and OPI scores. Many of the sub-variables seemed to have the same effect on the dependent variable. This redundancy in the analysis may result in multicollinearity. Multicollinearity in logistic regression models is the result of strong correlations between independent variables. The existence of multicollinearity may result in wrong signs and magnitudes of regression coefficient estimates, and consequently in incorrect conclusions about relationships between independent and dependent variables.

Only two independent variables showed statistical significance ($p \leq .05$) in predicting the dependent variable of agreement between CDS and OPI. Education level had an effect in that students who earned an associate degree were more accurate in predicting their speaking proficiency level (p. = .047). OPI qualification at ILR level 2 or higher had an effect in that students who received Level 2 or higher on the OPI were more accurate in predicting their speaking proficiency level, which indicates that students with high level proficiency more likely to be accurate in their self-assessment (p. = .000) (see Table 23).

Table 23: Results of Variables in the Equation in the Full Model (*n* =350)

Step 7	B	S.E.	df	Sig.
Education			2	.069
Ed_ AA(1)	-.584	.294	1	**.047**
Ed_ BA(2)	.049	.425	1	.909
M. Branch			3	.334
M. Branch(1)	-.605	.368	1	.100
M. Branch(2)	-.247	.382	1	.517
M. Branch(3)	-.227	.432	1	.599
Heritage	-.597	.413	1	.149
OPI qualification L2	1.580	.245	1	**.000**
Constant	.660	.536	1	.218

Figure 6: Observed Groups and Predicted Probabilities

```
         80 +
+
            |                                                                       Y
|
            |                                                                       Y
|
F           |                                                                       Y
|
R        60 +                                                                       Y
+
E           |                                                                       Y
|
Q           |                                                                       Y
|
U           |                                                                       Y
|
E        40 +                                                               Y       Y
+
N           |                                                               Y       Y    Y
|
C           |                                                               Y       Y    Y
|
Y           |                       Y                                       Y       N   YY      Y
|
         20 +                       N      Y     Y                          Y       N   YY      Y
+
            |                       N      N    YN       Y                  Y       N   YY      Y
|
            |                       N      NY   NN       YY     Y           N       N   YN      Y  Y
|
            |                       N      NN   NN       NN     NY          N       N   NN      N  Y
|
Predicted
---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:  0        .1        .2        .3        .4        .5        .6        .7        .8        .9
1
  Group:
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY

          Predicted Probability is of Membership for YES
          The Cut Value is .50
          Symbols: N - NO
                   Y - YES
          Each Symbol Represents 5 Cases.
```

Simultaneous Model of the Reduced Sample CM/KP Students

The purpose of using the stepwise model was to eliminate independent variables that had the minimum level of effect in predicting agreement between CDS and OPI. An additional purpose was to identify the variables that likely indicated the highest significance in predicting perfect agreement between CDS and OPI scores. Since logistic regression calculates the probability of success over the probability of failure, the results of the analysis are in the form of an odds ratio (see Figure 6 above). The simultaneous model of logistic regression also provides knowledge of the strength of relationships among the independent variables represented by the formula on page. 120.

Based on the outcome of the previous backward stepwise regression model of the analysis, step 7 (see Table 24) was selected based on the highest level of accuracy (77.7%), in predicting agreement between CDS and OPI scores for Chinese Mandarin and Korean students.

An additional variable (*attended in-country immersion*) was added to the simultaneous model since some Chinese and Korean students had the opportunity to travel to South Korea and China for four to six weeks to receive language instruction and practice the language in the target country. It was hypothesized that this real-life, in-country experience would help improve the accuracy of self-assessments of speaking proficiency. Arabic students were excluded from this model since they did not have the opportunity to travel to an Arabic speaking country due to the Arab revolutions in Egypt and Tunisia.

Although gender was eliminated in the full model due to its weakness in predicating perfect agreement, the researcher decided to include gender to detect the

possible effect of gender on CDS and OPI agreement. The reason for including gender was to examine if teachers' culture had an effect on the accuracy of self-assessment between Asian culture (CM-KP) and the Arabic culture. The level of accuracy in the stepwise model was (58.3%) and it increased to (77.7%) after including immersion and gender in the simultaneous model (see Tables 21 and 24).

Table 24: Simultaneous Model Level of Accuracy (CM-KP)

		Predicted		
		Perfect OPI/CSD Agreement		
Observed		NO	YES	Percentage Correct
Perfect OPI/CSD Agreement	NO	46	16	74.2
	YES	13	55	80.9
	Overall Percentage			77.7

Table 24 presents the level of accuracy of the model in predicting exact agreement between OPI and CDS. Based on the figures in the table there was a total of 62 observed discrepancies between OPI and CDS scores in CM and KP (the "NO" row). 46 of the 62 discrepancies (74.2%) were predicted by the model. Similarly 55 of 68 observed agreements (80.9%) in (the "YES" row) were predicted by the model. The overall percentage correct was calculated as follows: 46/62 + 55/68 = 101/130 = 77.7%. The simultaneous model confirmed the use of a simplified linear model, and it also gave clearer insight in determining the key factors that predict the dependent variables. Table 25 shows the variables that were tested in the simultaneous model. The variables evaluated were Military Branch, Education, Heritage Speaker, OPI Qualification at ILR=2, Type of Language, Gender, and Attended In- country Immersion.

Table 25: Categorical Variables in the Simultaneous Model (CM-KP) (*n*=130)

		Frequency	Parameter coding		
			(1)	(2)	(3)
Military Branch	ARMY	41	1.000	.000	.000
	AIR FORCE	48	.000	1.000	.000
	MARINES	18	.000	.000	1.000
	NAVY	23	.000	.000	.000
Education	GED/High School	84	1.000	.000	
	Associate Degree	17	.000	1.000	
	College Grad	29	.000	.000	
Heritage Speakers	NO	114	1.000		
	YES	16	.000		
OPI Qualification at ILR=2	Not Qualified	64	.000		
	Qualified	66	1.000		
Type of Language	CM	65	.000		
	KP	65	1.000		
Gender	Female	32	.000		
	Male	98	1.000		
Attended In- country Immersion	NO	99	.000		
	YES	31	1.000		

The results of the regression equations in Table 26 include the unstandardized model coefficients (B), the associated standard errors (SE), and the significance values for the predictor variables. The results indicated that students in the Army branch were more accurate in predicting their speaking proficiency level (p. = .029). Results showed that male were more accurate in predicting their speaking proficiency level (p. = .014).

Students who received level 2 or higher in the final OPI were more accurate in predicting their speaking proficiency level (p. = .000). These independent variables showed the highest association with agreement between CDS and OPI.

Table 26: Results of Variables in the Simultaneous Model (CM/KP) (n=130)

Type of Variable	B	S.E.	df	Sig.
Education			2	.907
Ed_ MA(1)	-.190	.550	1	.730
Ed_ BA(2)	.044	.752	1	.953
M. Branch			3	.132
M. Branch(1)	-1.548	.711	1	**.029**
M. Branch(2)	-.557	.666	1	.404
M. Branch(3)	-.975	.857	1	.255
Heritage (1)	-1.155	.677	1	.088
Immersion(1)	-.721	.574	1	.209
Language(1)	.370	.507	1	.465
Gender(1)	-1.364	.554	1	**.014**
OPI-Qual	-2.030	.474	1	**.000**
Constant	4.133	1.127	1	.000

The outcome of the joint predictive ability of all the covariates in the model was presented in the omnibus test (see Table 27). The Chi-square statistics are all the same because stepwise logistic regression or blocking was not used. The value given in the significance column is the probability of obtaining the chi-square statistics by chance. This is of course, the p-value, which is compared to the critical value, .05, to determine the overall model is statistically significant. In this study, The Chi-square test statistic is,

χ^2 = 47.622, with df = 10 and a p-value of 0.000. The model is statistically significant because the p. value is less than .000. Therefore, the null Hypothesis can be rejected. (χ^2 (df, 10,130) = 47.6, p < .05).

Table 27: Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	47.622	10	.000
	Block	47.622	10	.000
	Model	47.622	10	.000

Table 28, the Model Summary, reports the R-Square estimated value, $R^{2 =}$ .409, which indicates that the logistic regression model fit the data. This measure is therefore useful when comparing several different logistic regression models. The Cox & Snell R Square and the Nagelkerke R Square are pseudo R-squares. There are a wide variety of pseudo-R-square statistics (these are only two of them). This statistic does not mean what R-squared means in regular regression (the proportion of variance explained by the predictors), this statistic should be interpreted with great caution.

Table 28: Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	132.320[a]	.307	.409

Summary

In Chapter 4, findings from the study were presented. Research Hypothesis 1 stated that a relationship exists between students' scores on a self-assessment survey and those on a criterion- referenced speaking proficiency test. Research Hypothesis 1 was supported.

Research Hypothesis 2 stated there is an impact of various variables on the psychometric properties (validity, reliability and accuracy) of student self-assessments of second language speaking proficiency. The results indicated that a group of the independent variables, or "constellations" of Military Branch, Gender and The OPI Qualification at ILR Level 2, showed the greatest association with the agreement between CDS and OPI. Thus, Null Hypothesis 2 can be rejected.

CHAPTER 5

CONCLUSIONS, IMPLICATIONS, AND RECOMMENDATIONS

This final chapter consists of three sections. The first section will summarize the study and its conclusions. The second section will discuss the implications of the study for foreign language assessment and classroom practice. The final section will provide recommendations for future research in the study of self-assessment in foreign languages.

Summary of the Study

The primary purpose of the study was to develop and validate a self-assessment instrument that can be used to obtain a highly reliable estimate of the foreign language proficiency of native speakers of English. A secondary purpose of the study was to determine whether selected variables (i.e., type of foreign language, education level, military branch, military rank, gender, age, immersion in the target country, heritage-speaker status, prior training in foreign languages, and qualification at Level two or higher on the OPI) might affect student's self-assessment in speaking foreign languages.

The study took place at the Defense language Institute Foreign Language Center (DLIFLC) in Monterey, California. The study involved five schools teaching foreign languages--three teaching Arabic, one teaching Chinese, and one Korean. The sample was composed of students (n=350) who were members of four military branches: Army, Navy, Marines, and Air Force. The sample was drawn from the basic-course programs--220 students participated in Arabic, 65 in Chinese, and 65 in Korean.

The study was composed of four phases: 1) validating the Can-Do- Scale (CDS) survey which was specifically designed for this study , in terms of its content and

construct validity, 2) establishing the test-retest reliability of the two forms of the CDS, with the forms being administered approximately five to ten days apart, 3) conducting a pilot study of the relationship between the CDS and OPI in five Category-III languages, Persian Farsi, Persian Dari, Tagalog, Thai, and Urdu, and 4) administering the CDS a week to ten days prior to the final end-of course OPI.

Conclusions

Hypothesis 1

H1: There is a relationship between student scores on a self-assessment survey and those on a criterion-referenced speaking proficiency test.

Before the correlation was computed between the CDS and OPI scores, the following steps were conducted: 1) a test-retest reliability analysis on different forms of the CDS, 2) examine the inter-rater reliability of both the CDS and the OPI.

The test-retest reliability study was conducted with students studying Arabic in semesters I, II, and III. The results of this study indicated that the CDS was significantly reliable in all three semesters, in semester I Spearman's rho was $r = .638$, $p < .05$ in semester II it was $r = .902$, $p < .05$ and in semester III it was $r = .803$, $p < .05$.

The inter-rater reliability of the two raters of the CDSs was statistically significant in the entire sample (n=350) of Arabic, Chinese and Korean students. The Spearman's rho correlation between the original, independent ratings of the two testers who scored the CDSs was ($r = .943$, $p < .05$), which indicated that the CDS could be scored reliably.

The inter-rater reliability of the OPI testers who rated independently was extremely high also (r = .975, p < .05). When the two OPI ratings for a given student differed, the majority typically varied by only a plus level. Third ratings were conducted independently by single master testers. For quality-control purposes, tape recordings of 164 (47%) of the 350 OPIs conducted in the study were third-rated by master testers. In 145 (88.4%) of the 164 third-rated tests, the third raters agreed with the initial raters. In this study, there were a few cases where the third rater disagreed with the initial OPI raters. There were cases where fourth raters felt the OPI was not given the correct rating and they recommended a new OPI. These cases occurred in Arabic in eleven tests , two test in Korean, and two tests in Chinese.

The study showed that students' self-assessment scores correlated positively with their final OPI ratings, and that the correlation was statistically significant at the p = .05 level. Students' self-assessment scores, as measured by the (CDS), correlated with their final OPI scores (r = .272, p < .05). Although the correlation was statistically significant, it was low, and indicated that the null hypothesis of no relationship between CDS and OPI scores could be rejected.

A simple percent agreement statistic was calculated to determine the percentage of agreement between CDS and OPI. The percentage of perfect agreement between CDS and OPI scores was moderate: 204 out of the 350 students, or 58.3%, received the exact same score on both measures. In the case of the remaining students, who had discrepant ratings, the majority of the ratings were only a plus level apart.

The highest number of cases of perfect agreement between CDS and OPI scores occurred at Level 1+. Out of 215 students who received Level 1+ on the final OPI, 154

students (71.6%), rated their proficiency at Level 1+ on the CDS. The second highest level of agreement was at Level 2. 131 students received Level 2 on the final OPI, while 48 of them (36.6%), assessed their proficiency at Level 2 on the CDS.

For the students who received different scores on both measures, one is led to ask how far their CDS score was from their OPI score. The value of this study is to determine how well students self-assess their foreign language: Overestimation, Exact Rating, and Underestimation of their speaking proficiency.

As shown in table 29, among students who were studying Arabic, 136 (61.8%) of the 220 students received the same score on both measures. Arabic students received the highest percentage of agreement between CDS and OPI within the languages examined in this study (61.8%).

Table 29: Exact Agreement, Over or Under-Rating in Arabic (n=220)

			Survey Final Rating						
			0+	**1**	**1+**	**2**	**2+**	**3**	**Total**
OPI Final Rating	1+	Count	1	2	105	25	18	0	151
		% within OPI Final Rating	.7%	1.3%	69.5%	16.6%	11.9%	.0%	100.0%
		% of Total	.5%	.9%	47.7%	11.4%	8.2%	.0%	68.6%
	2	Count	0	0	27	30	10	1	68
		% within OPI Final Rating	.0%	.0%	39.7%	44.1%	14.7%	1.5%	100.0%
		% of Total	.0%	.0%	12.3%	13.6%	4.5%	.5%	30.9%
	2+	Count	0	0	0	0	1	0	1
		% within OPI Final Rating	.0%	.0%	.0%	.0%	100.0%	.0%	100.0%
		% of Total	.0%	.0%	.0%	.0%	.5%	.0%	.5%
	Total	Count	1	2	132	55	29	1	220
		% within OPI Final Rating	.5%	.9%	60.0%	25.0%	13.2%	.5%	100.0%
		% of Total	.5%	.9%	60.0%	25.0%	13.2%	.5%	100.0%

Level 0+, 1 and 3 were excluded from this analysis because there were so few students with CDS scores at these levels, that the percent agreement statistics would not be meaningful. Levels 1+, 2 and 2+ were examined for over or under-rating for students who did not receive the same score on CDS and OPI. Table 29 indicates that 132 students

received Level 1+ on the CDS; while 27 (20.5%) of them underestimated their proficiency at Level 2 on the OPI measure.

In regards to overestimating, among students who were studying Arabic, at Level 2, 55 students received a 2 on the CDS while 25 of them (45.5%) overestimated their proficiency at level 1+ on the final OPI. At Level 2+, 29 students received a 2+ on the CDS , while 28 of them or 96.7% overestimated their proficiency by at least a plus level, including 18 (62.1%) who overestimated their speaking proficiency by a full level (see Table 29).

As shown in table 30, among students who were studying Korean, 36 (56.9%) of the 65 students received the same score on both measures. Korean students received the second highest percentage of agreement between the CDS and OPI scores within the languages examined in this study (56.9%).

Level 1 and 3 were excluded from this analysis because there were so few students with CDS scores of these levels, thus the percent agreement statistics would not be meaningful. In regards to underestimating, Table 30 indicates that 49 students received Level 1+ on the CDS; while 17 (34.7%) of them underestimated their proficiency at Level 2 on the OPI measure.

Table 30: Exact Agreement, Over or Under-Rating in KP (n=65)

OPI Final Rating		Survey Final Rating					
		1	1+	2	2+	3	Total
1+	Count	1	32	6	1	0	40
	% within OPI Final Rating	2.5%	80.0%	15.0%	2.5%	.0%	100.0%
	% within Survey Final Rating	100.0%	65.3%	60.0%	25.0%	.0%	61.5%
2	Count	0	17	4	2	0	23
	% within OPI Final Rating	.0%	73.9%	17.4%	8.7%	.0%	100.0%
	% within Survey Final Rating	.0%	34.7%	40.0%	50.0%	.0%	35.4%
2+	Count	0	0	0	1	1	2
	% within OPI Final Rating	.0%	.0%	.0%	50.0%	50.0%	100.0%
	% within Survey Final Rating	.0%	.0%	.0%	25.0%	100.0%	3.1%
Total	Count	1	49	10	4	1	65
	% within OPI Final Rating	1.5%	75.4%	15.4%	6.2%	1.5%	100.0%
	% within Survey Final Rating	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

In regards to overestimating among students who were studying Korean, at Level 2, 10 students received Level 2 on the CDS , while 6 of them (60.0%) overestimated their proficiency at level 1+ on the final OPI. At Level 2+, 4 students received a 2+ on the CDS, while one of them overestimated their proficiency by at least a full level, two students overestimated by a plus level. Three (75.0%) of the four students who received 2+ on the CDS overestimated his/her speaking proficiency by at least a plus level (see Table 30).

Table 31: Exact Agreement, Over or Under-Rating in CM (n=65)

			Survey Final Rating				
			1+	2	2+	3	Total
OPI Final Rating	1+	Count	17	3	4	0	24
		% within OPI Final Rating	70.8%	12.5%	16.7%	.0%	100.0%
		% of Total	26.2%	4.6%	6.2%	.0%	36.9%
	2	Count	19	14	5	2	40
		% within OPI Final Rating	47.5%	35.0%	12.5%	5.0%	100.0%
		% of Total	29.2%	21.6%	7.7%	3.1%	61.5%
	2+	Count	0	1	.0%	0	1
		% within OPI Final Rating	.0%	100.0%	.0%	.0%	100%
		% of Total	.0%	1.5%		.0%	1.5%
	Total	Count	36	18	9	2	65
		% within OPI Final Rating	55.4%	27.7%	13.8%	3.1%	100.0%
		% of Total	55.4%	27.7%	13.8%	3.1%	100.0%

As shown in table 31, among students who were studying Chinese, 31 (47.7%) of the 65 students received the same score on both measures. Chinese students received the lowest percentage of agreement between the CDS and OPI scores within the languages examined in this study (47.7%).

Level 3 was excluded from this analysis because there were so few students with CDS scores of this level, thus the percent agreement statistics would not be meaningful. In regards to underestimating, Table 31 indicates that 36 students received Level 1+ on the CDS; while 19 (52.8%) of them underestimated their proficiency at Level 2 on the OPI measure.

In regards to overestimating, among students were studying Chinese, at Level 2, 18 students received a 2 on the CDS , while 3 of them (16.7%) overestimated their proficiency at level 1+ on the final OPI. At Level 2+, 9 students received a 2+ on the CDS and none of them received a 2+ on the OPI measure. Five (55.6%) of the 9 students overestimated their proficiency by a plus level, while 4 (44.4%) of them overestimated their speaking proficiency by a full level (see Table 31).

Table 32: Language Summary AD-CM-KP (n=350)

Language	Exact Agreement	Overestimating	Underestimating
Arabic	136/220 = 61.8%	54/220 = 24.5%	30/220 = 13.6%
Korean	37/65 = 56.9%	10/65 = 15.4%	18/65 = 27.7%
Chinese	31/65 = 47.7 %	14/65 = 21.5%	20/65 = 30.8%
Entire Sample	204/350 = 58.3	78/350 =22.3%	68/350 = 19.4%

There was sufficient evidence from the findings of the study to conclude that students in the Asian languages underestimated their ability to speak their foreign language, as measured by the ILR scale (see Table 32). That might be due to an association between teachers' social norms which is modesty in self-rating. In other words, Asian teachers do not over praise students' performance. Teacher attempts to create learning environments by using praise may actually be counterproductive. Brophy, (1981) Children have an intrinsic desire to learn new language by nature. Ineffective praise can stifle students' natural curiosity and desire to learn by focusing their attention on extrinsic rewards rather than the intrinsic rewards that come from the task itself.

The highest percentage of students who underrated their proficiency was CM (30.8%), the second highest was KP (27.7%), and the lowest was AD (13.6%). The highest percentage of students who overrated their proficiency was AD (24.5%), the second highest was CM (21.5%), and the lowest was KP (13.6%) (see Table 32).

Hypothesis 2

H2: There is an impact of various variables on the psychometric properties (validity, reliability and accuracy of student self-assessments of second language speaking proficiency.

Three independent variables were significant predictors of exact agreement between CDS and OPI scores:

1) The Qualification of OPI at ILR Level 2: members of the sample who had a better command of their target language (as indicated by a score of Level 2 or higher on the official OPI) were more accurate in their self-assessment;

2) Gender: males self-assessed their speaking proficiency in their foreign language more accurately than females.

3) Military Branch: out of the four military branches that participated in the study, Army, Air Force, Marines, and Navy, students serving in the Army evaluated their speaking proficiency in foreign languages more accurately than students in the other military branches. One could hypothesize that might be due to the strict discipline and emphasis on accuracy that the Army imposes on its service members, based on my personal experience and the feedback I have received from classroom instructors

The predictors of The Qualification of OPI at ILR Level 2, Gender, and Military Branch of the Army were significant for the outcome of perfect agreement, thus Null

Hypothesis 2 can be rejected. At least one of the independent variables significantly predicts the dependent variable outcome of perfect agreement.

Implications for the Profession

Implications for Foreign Language Assessment and Instruction

The main challenge in this study was to make sure that the instrument, the Can-Do Scale (CDS) can be trusted to provide an accurate estimate of people's proficiency. This study found that there was sufficient evidence that the self-assessment survey is reasonably valid, reliable and yields reasonably accurate ratings of students' speaking proficiency. Although the correlation between CDS and OPI scores was statistically significant, it was low (r = .272). The overall level of perfect agreement was 58.3%: 204 out of the 350 subjects who participated in the study received the same rating on both measures.

146 of the 350 subjects (41.7%) did not obtain the same rating. For 119 (34.0%) out of the 350 subjects the ratings on the CDS and OPI only differed by +/- a plus level. Only 27 of the 350 participants (7.7%) had CDS and OPI scores that differed by more than a plus level. The 27 all overestimated their level of proficiency

In conclusion, 323 (92.3%) out of the 350 subjects had CDS scores that were either exactly the same as their OPI score or only a plus level above or below it. Hence, the CDS was able to predict a fairly close rating of 92.3% of the subjects who participated in the study.

DLIFLC is the main institution that provides proficiency testing for foreign languages for the Department of Defense (DoD) and is unable to satisfy the increasing demand for testing. DLIFLC does not have an adequate number of testers to fulfill the

high number of testing requests received since the 9/11 attack on the USA. Thus, the CDS could be used to meet this need. The speaking ability of military personnel in the field could be assessed no matter where they are located, especially when it is difficult to schedule a face-to-face or telephonic test.

The CDS could be used to evaluate the speaking proficiency of soldiers in war zones. Commanders in the field could have an up- to-date estimate of what their soldiers can do in real life. When commanders know the approximate ability of military linguists on hand, they can assign linguists to do tasks that are compatible with their ability and thus possibly save lives in wartime. Commanders would of course need to take into consideration the fact that the soldiers' CDS scores might be higher or lower by a plus level than their true proficiency scores.

This study has shown that students are able to evaluate their speaking proficiency in foreign languages with some degree of accuracy. The self-assessment survey could therefore be used as an alternative form of assessment in classrooms at the end of each semester to measure students' progress. DLIFLC Basic Course language programs consist of three semesters. Students take an achievement speaking test at the end of each semester and most of the time the test is limited in content, which raises questions about its validity as a measure of proficiency. Yet, there is of course nothing inherently wrong with a speaking *achievement* test. It depends on the purpose of the test. The CDS covers proficiency levels 0+ to 3, and it is less stressful than a formal speaking test that may cause students some embarrassment if or when they make errors. The use of self-assessment can help students to estimate their genuine proficiency in a less-stressful learning environment.

The self-assessment instrument, or the CDS, may be used as a Diagnostic Assessment (DA) tool that is less stressful in nature than an oral interview or a regular DA, because no passing or failing grade is given, just a proficiency level. Diagnostic assessment is a formative approach that can be used to assess students' progress in the development of speaking, listening, and reading skills. The advantage of using the CDS as DA tool is that students would be able to identify their strengths and weaknesses in speaking proficiency compared to the ILR scale. The CDS could thus be used as a pedagogical assessment tool to enable learners to advance in using the target language.

The purpose of using the CDS in DA evaluation is to provide effective guidance for improving proficiency in ways that can be tracked and measured by the ILR scale. The CDS can give students the chance to evaluate their speaking skills and enhance their self-confidence since they take part in the evaluation process with no fear of failing. Diagnostic assessment can be used as a formative assessment tool for the purpose of guiding the acquisition of foreign language proficiency in a more efficient and effective manner than is normally experienced in a typical contemporary classroom (Fahmy, 2007). The CDS could be used as an alternative to formal proficiency testing to measure how well a learner can perform real-life speaking tasks. The CDS can be used as a DA tool to alleviate the pathway for each learner toward higher proficiency through a tailored curriculum or a personalized plan for study.

DLIFLC's missions are teaching, testing, and sustainment for Department of Defense (DoD) personnel to ensure the safety of the nation. The DLIFLC's Directorate of Continuing Education (CE) provides post-basic education through resident and non-resident programs. The school of continuous education provides intermediate and

advanced courses. The CDS can be a valid assessment test to be used for placement purposes in intermediate and advanced program. Using the CDS in the placement process could save a great deal of financial resources by eliminating formal, face-to face OPI tests. The test could provide instructors with a holistic perception of the students' proficiency in speaking; thus instructors could tailor a program to a particular group of students sharing similar needs. Program leaders could benefit from what they learn from the test results to select appropriate materials and facilitation techniques that would augment the success of the program.

The findings of this study indicate that the CDS is reasonably reliable in estimating students' proficiency in speaking. This author suggests that DLIFLC consider using the CDS in its education programs. The CDS could have multiple implications for teaching and assessing foreign languages in the following areas:

1) It can be used as an alternative way of assessing speaking for military linguists who are located in the US or abroad.

2) It can be used as a formative assessment tool for the basic course, intermediate and advanced programs.

3) It can be used as a placement test to identify the proficiency level of a linguist before he or she enrolls in the resident and non-resident programs.

4) The CDS can be used as a diagnostic assessment tool to measure students' progress in the development of speaking proficiency and to offer them practical guidance.

"Without self-assessment there is no self-awareness, and self-awareness is important to language learning, both in terms of knowing what level one has already

achieved, and in terms of knowing one's strengths and weaknesses, and one's preferred way of learning and developing" (Alderson, 2004, P. 243).

One of the challenges that students and faculty at DLIFLC have is understanding the jargon of the ILR scale. ILR is the core of every academic activity at DLIFLC: curriculum development, test development, faculty development and instruction. Lack of comprehension of the scale affects the achievement of DLIFLC goals. Using the CDS as a diagnostic assessment tool will enhance comprehension of the ILR criterion among teachers and students, since CDS is an ILR-based instrument.

Improving teachers' knowledge of the ILR scale would enhance their ability to understand the ILR descriptors and provide students with guidance to reach a specific proficiency level. Additionally, learners would increase their knowledge of what they need to do to achieve desired proficiency levels. Finally, implementing self-assessment in foreign language education in general would increase learners' knowledge of their learning goals and their learning needs, and thus might enhance their motivation and goal orientation.

DLIFLC students studying category IV languages have expressed concern about the length of the language program--63 weeks. These programs are highly intensive and students' learning deteriorates toward the end of the course. Using the CDS toward the end of the course could increase students' self-confidence in their speaking ability, help them identify any weaknesses, and motivate them to improve.

Recommendations for Future Research

The study focused on a very specific military population at DLIFLC. Future research could be conducted in non-military educational institutions and less-restricted environments. The research question regarding this aspect could be the following: 1) what role does educational environment (military-civilian) play in the accuracy of a student's self-assessment of his/her own proficiency in speaking?

In addition, the study focused on students who were studying Arabic, Chinese, and Korean, which are Category IV languages, i.e., languages that are considered the most difficult languages for native speakers of English. Further research should consider targeting less difficult languages (Category I, II and III). These types of studies would focus on the question of whether a language's level of difficulty for native speakers of English can affect students' self-assessment of proficiency in that language.

The length of a Category IV language course at DLIFLC is 63 weeks; thus examining the length of the course--24 weeks in category I; 36 weeks in category II; and 47 weeks in category III might have an effect of the outcome of self-assessment in foreign languages.

1) Is there a relationship between the levels of language difficulty for native speakers of English and the accuracy of self-assessment in speaking?

2) What role does the length of the language program play in the accuracy of a student's self-assessment of his/her own proficiency in speaking?

Lundsteen, (1979) Listening and reading comprehension, are usually defined as a receptive skills comprising both a physical process and an interpretive, analytical process. The expanded definition of listening also emphasizes the relationship between listening

and speaking. The role of comprehensible input has been prominent in second language acquisition (SLA) research and theory, particularly in the past twenty years. This has been driven by the belief that a learner's exposure to the target language is not in itself a sufficient condition for second language (L2) acquisition. From Corder's (1967) early claims of input and intake to Krashen's (1982) Input Hypothesis and Long's (1983) Interaction Hypothesis, there has been a common conviction that input must be comprehended by the learner if it is to assist the acquisition process. Future researchers should consider a study on the relationship between speaking self-assessment and level of proficiency in reading and listening. The research questions involved could be the following:

1) Is there a correlation between learners' accuracy in self-assessment in speaking and level of proficiency in listening?

2) Is there a correlation between learners' accuracy in self-assessment in speaking and level of proficiency in reading?

One of the areas that concern many faculty members at DLIFLC deals with the students' motivation to learn specific languages. Military students often do not have the option to select the language that they desire to learn. Students' placement in different languages is entirely based on their DLAB scores and the needs of their military service, not on their particular need and desire. Motivation has been identified as the learner's orientation with regard to the goal of learning a second language (Crookes and Schmidt, 1991). Based on the outcome of my study, it would be interesting to find whether students' motivation affects the accuracy of their self-assessments. According to Blanche and Merino (1989), research in the area of attitudes and behavior as they relate to self-

assessment has been limited in scope, dealing frequently with the effects of attitude on achievement in second and foreign languages.

The research questions could be the following:

1) What role does students' motivation play in the accuracy of a student's self-assessment of his/her own proficiency?

2) Are certain personality types more accurate in their self-assessments of their own proficiency than others (DeMent, L. 2008)?

3) Does level of aptitude have an effect on the students' ability to assess their own speaking proficiency accurately?

Summary

This study developed and validated a self-assessment instrument which was a Can-Do Scale (CDS) that could be used to obtain a reliable estimate of the foreign language proficiency of native speakers of English. To realize this purpose, the researcher followed the following steps:

1) Investigated the relationship between two types of measures of oral proficiency: level scores inferred from a self-assessment instrument (the CDS) and ratings obtained from a formal Oral Proficiency Interview (OPI).

2) Investigated the impact of various variables that the literatures suggested were likely to affect the validity, reliability and the accuracy of scores and how well students assessed their speaking proficiency in foreign languages.

In the second step the study investigated whether the following variables: Type of Foreign Language, Education Level, Military Branch, Military Rank, Gender, Age,

Attended Immersion in the Target Country, Heritage Speaker, Prior Training in Foreign Languages, and The Qualification at ILR Level 2 might affect student's self-assessment of speaking ability in foreign languages.

A correlational research method using Spearman's rho correlation coefficient was employed to test if there was a relationship between students 'self-assessment as measured by the CDS and the final OPI. Results showed there was a practically significant (r =.272) relationship between students' self-assessment and final OPI score. Although the correlation was low, the percentage of exact agreement between CDS and OPI was moderately high, 204 out of 350 students, or 58.3%, received the same rating on both measures. When students did not obtain the same ratings, most were not far off from each other: 119 students out of 350, or 34% of the entire sample, were off by +/- an ILR plus level. 27 out of the 350 students, or 7.7%, over-estimated their speaking by more than a plus level.

Results for the entire sample showed that students had the highest level of agreement between the CDS and OPI at level 1+. The second highest level of agreement was at Level 2. The results thus indicated that the CDS was significantly reliable in predicting the OPI score at level 1+ and 2.

The findings of the study showed that the students in the Asian languages, Chinese Mandarin and Korean, underestimated their proficiency, which might reflect the influence of teachers' cultural norms in the Asian languages. In Chinese, 20 of the 65 students (30.8%) received a higher score on the OPI than the CDS. Thus almost (31.0%) of the students of Chinese underestimated their speaking proficiency (Table 32 on p. 144)

Usually the Chinese people do not like to show a high opinion of their own qualities. “Instead they are always modest about their successes, or prefer a low-key statement to a display of their advantages. When you praise a Chinese person, he humbly tells you how deficient he is”. http://www.echineselearning.com.

Among the students who were studying Korean, 18 of the 65 (27.7%) received a higher score on the OPI the CDS. Thus almost (28%) of the students of Korean underestimated their speaking proficiency (see Table 32 on p. 140).

“Koreans consider modesty very important, especially in interactions with people to whom they need to show respect. While in the West, honesty is very important, in Korea, modesty is more important than honesty.” http://www.emagasia.com.

“Koreans know as well as others that they in reality can do it well, but need to demonstrate a sense of modesty at all the times.” On the other hand, in a job interview or similar situation, unpracticed at bragging, Koreans can sound as if they are incompetent. http://www.emagasia.com.

In contrast, in the Arabic language the majority of students who did not get the same rating on both measures over-estimated their speaking proficiency. Fifty four (24.5%) out of the 220 students who participated in the Arabic study over estimated their speaking ability by at least a plus level. The over rating may be due to the influence of the culture of their Arabic teachers. In the researcher’s opinion, Arabic teachers at the DLIFLC tend to over praise their students’ ability, rather than giving constructive feedback and analyzing strengths and weaknesses.

A logistic regression analysis was conducted to assess the association between the dependent variable of the level of agreement between CDS and OPI ratings, and the

following ten independent predictor variables: Type of Foreign Language, Education Level, Military Branch, Military Rank, Gender, Age, Attended Immersion in the Target Country, Heritage Speaker, Prior Training in Foreign Languages, and The Qualification at ILR Level 2. Results showed that a constellation of three variables, Military Branch, Gender, and The OPI Qualification at ILR Level 2, showed the highest association with the level of agreement between CDS and OPI scores.

REFERENCES

Adams, Marianne L. (1980). Five co-occurring factors in speaking proficiency. In James R. Frith (Ed.), *Measuring spoken language proficiency* (pp. 1–6). Washington, DC: Georgetown University Press.

Adams, T. L. (1998). Alternative assessment in elementary school mathematics, *Childhood Education*, (4), 220–224.

Agresti, A. (1996). An Introduction to Categorical Data Analysis. John Wiley and Sons, Inc.

Alderson, J. C. (2004). *Diagnosing Foreign Language Proficiency the interface between learning and assessment.* London: Continuum

Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership, 57* (5), 13-18.

Andrade, H. (2001). The effects of instructional rubrics on learning to write. *Current Issues in Education, 4*(4).

Andrade, H. & Boulay, B. (2003) Gender and the role of rubric-referenced self-assessment in learning to write, *Journal of Educational Research,* (1), 21–34.

Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education, 32*(2), 159-181.

Andrade, H., & Valtcheva, A. (2009). Promoting Learning and Achievement through Self-Assessment. *Theory into Practice, 48*(1), 12-19.

Andrade, H. L., Wang, X., Du, Y., & Akawi, R. L. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *Journal of Educational Research, 102*(4), 287-302.

Armstrong, T. (1994). Multiple Intelligences in the Classroom. *Association for Supervision and Curriculum Development*. Alexandria, Virginia

Arter, J., & Chappuis, J. (2007). Creating and recognizing quality rubrics. Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin Press.

Au, K.H. (1993). Literacy instruction in multicultural settings. Orlando: Harcourt Brace Jovanovich.

Bachman, L. (1990). *Fundamental consideration in language testing.* Oxford: Oxford University Press

Bachman, L. F. (2000). Learner-directed assessment in ESL. In: G. Ekbatani & H. Pierson (Eds.), *Learner-directed assessment in ESL* (pp. ix-xii). New Jersey: Lawrence Erlbaum Associates, Inc.

Bachman, L. & Palmer, A. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing*, *6*(1), 14-29.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Badolato, L.A. (1998). Recognizing and meeting the special needs of gifted females. Gifted Child Today, *21*(6), 32-37.

Bailey, K. & Lazar, J. (1976). Accuracy of self-rating of intelligence as a function of sex and level of ability in college students. *Journal of Genetic Psychology*, 129-279-290.

Bandura, A. (1969). *Social learning theory*. New York: General Learning Press.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84, 191.215.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist,37,* 122-147.

Bandura, A. (1986). *Social foundations of thought and action.* Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. Educational Psychology, *28*(2), 117-148.

Barcinas, J. (2011). Private conversation. Test Management, DLIFLC, Monterey, CA.

Belasco, A. (2011). The Cost of Iraq, Afghanistan, and Other Global War on Terror Operations since 9/11. Congressional Research Services.

Bell, B. S., & Ford, J. K. 2007. Reactions to skill assessment: The forgotten factor in explaining motivation to learn. *Human Resource Development Quarterly, 18,* 33–62.

Blanche, P. (1988). Self-assessment of foreign language skills: Implications for teachers and researchers. *RELC Journal, 19,* 75-96.

Blanche, P., & Merino, B. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning, 39,* 313-340.

Boud, D. (1986). *Implementing student self-assessment.* Sydney: HERDSA.

Boud, D. (1991). *Implementing student self-assessment*. Sydney, Australia: Higher Education Research and Development Society of Australia.

Boud, D. (1995). *Enhancing learning through self-assessment.* London: Kogan Page.

Boud, D. & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: a Critical Analysis of Findings. *Higher Education, 18* (5), 529-549.

Bourke, R. & Poskitt, J. (1997). *Self-assessment in the New Zealand classroom* (Booklet). Wellington: Ministry of Education.

Breidert, T. (2009) "Self-Assessments by U.S. Army Officers: Effects of Skill Level and Item Ambiguity on Accuracy." Masters Theses & Specialist Projects.

Brophy, J.E. (1981). "*Teacher Praise: A Functional Analysis*." REVIEW OF EDUCATIONAL RESEARCH 51(1) 5-32.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32,* 653-675.

Bruner, J. (1960). *The process of education.* Cambridge, MA: Harvard University Press.

Bruner, J. (1966). *Toward a theory of instruction.* Cambridge, MA: Harvard University Press.

Bruner, J. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.

Burns, A. (1999). *Collaborative action research for English language teachers.* Cambridge, England: Cambridge University Press.

Butler, Y. G. & Lee, J. (2006). On-task versus off-task self-assessment among Korean elementary school students studying English. *The Modern Language Journal, 90* (4), 506-518.

Calfee, R., & Hiebert, E. (1990). Classroom assessment of reading. In: R. Barr, M. Kamil, P. Mosenthal, & P.D. Pearson (Eds.), *Handbook of reading research (2nd ed.)*. New York: Longman, 281-309.

Campbell, D. J., & Lee, C. (1988). Self-appraisal in performance evaluation: Development versus evaluation. *Academy of Management Review,* 13, 302–314.

Carlson, et al, (1990). Experience Related to the Sojourn of Study Abroad Students and Changes in their Language Proficiency. In Jerry S. Carlson (Ed.), *Study Abroad: The Experience of American Undergraduates* (pp.). New York: Greenwood Press.

Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. *Innovations in Education and Teaching International, 44*(1), 57-66.

Cason, G. J., & Cason, C. L. (1984). A determine theory of clinical performance rating. *Evaluation and the health profession, 7,* 221-241.

Cohen, L., Manion, L., & Morrison, K. (2004). *A Guide to Teaching Practice* (5th ed.). London: Routledge Falmer.

Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics, 5,161-170*

Crichton N.J. (1999) Information point: Spearman's rank correlation. Journal of Clinical Nursing 8, 763.

Crookes, G., & Schmidt R.W. (1991). Motivation : Reopening the research agenda. *Language Learning*, 41(4), 469-512.

Deakin-Crick, R., Sebba, J., Harlen, W., Guoxing, Y., & Lawson, H. (2005). Systematic review of research evidence of the impact on students of self- and peer-assessment. *Protocol in Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

DeMent, L. (2008). *The relationship of self-evaluation, writing ability, and attitudes toward writing among gifted grade 7 language arts students.* Retrieved from *ProQuest Dissertations and Theses*.

De Saint Leger, D. (2009). Self-assessment of speaking skills and participation in a foreign language class. *Foreign Language Annals, 42*(1), 158-178.

Diaz, E. I. (1998). Perceived factors influencing the academic underachievement of talented students of Puerto Rican descent. *Gifted Child Quarterly, 42,* 105-122.

Dickinson, L. (1987). *Self-instruction in language learning.* Cambridge, England: Cambridge University Press.

Dickinson, L. (1992). Learner autonomy. Dublin: Authentik.

DLIFLC (2011). Defense Language Institute Foreign Language Center, Presidio of Monterey. *www.dliflc.edu*.

Dobransky, N. D., & Frymier, A. B. 2004. Developing teacher student relationships through out of class communication. *Communication Quarterly,* 52: 211–223.

Edwards, R. (1989). An Experiment in student self-assessment. *British Journal of Education Technology*, 20 (1), 5-10.

Ekbatani, G. & Pierson, H. (1998). *Teacher portfolios-vehicles of faculty assessment, reflection and growth*. (Eric Document Reproduction Service No. ED 163-753).

Ellis, R. (1994). *The study of second language acquisition.* Oxford: Oxford University Press.

Erwin, T. (1991). *Assessing student learning and development.* San Francisco: Jossey-Bass.

Evaluating Foreign Language Immersion Training, (1997). The Defense Language Institute Foreign Language Center, *Program Evaluation Research and Testing*.

Evans, K. A. (2001). Rethinking self-assessment as a tool for response. Teaching English in the Two-Year College, 28, 293–301.

Fahmy, M. (2007). The Various Purposes of Diagnostic Assessment between Idealism and Realism. DLIFLC

Falchikov, P., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. Review of Educational Research, 59(4), 395-430.

Fink, L. D. (2003). *Creating significant learning experiences: An integrated approach to designing college courses*. San Francisco: Jossey-Bass.

Fischer, H. (2010). U.S. Military Casualty Statistics: Operation New Dawn, Operation Iraqi Freedom, and Operation Enduring Freedom. Congressional Research Services (CRS).

Flavell, John, H. (1979). Metacognition and cognitive Monitoring. *America Psychologist, 34*, (10), 906-911.

Flinders, D. J. (1992). In search of ethical guidance: Constructing a basis for dialogue. *Qualitative Studies in Education,* 5(2), 101-115.

Fox, S., & Dinur, Y. (1988). Validity of self-assessment: A field evaluation. *Personnel Psychology, 41*, 581-592.

Frederiksen, J. & Collins, A. (1989) A systems approach to educational testing, *Educational Researcher,* 18(9), 27–32.

Galloway, Vicki B. (1987). From defining to developing proficiency. A new look at the decision. In Heidi Byrnes (Ed.), *Defining and developing proficiency: Guidelines, implementations, and concepts* (pp. 25–73). Lincolnwood, IL: National Textbook Company.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Gardner, H. (1993). *Frames of mind: The theory of multiple intelligences* (2nd ed.) London: Fontana.

Gardner, H. (1995). *The unschooled mind: How children think and how schools should teach.* New York: Basic Books.

Gardner, H. (1999). Intelligence reframed: Multiple intelligences for the 21st century. New York: Basic Books.

Gardner, H., & Hatch, T. (1989). Multiple intelligences go to school: Educational implications of the theory of multiple intelligences. *Educational Researche,*18(8), 4-9.

Gardner, R. C, & MacIntyre, P. (1991). Motivational variables in second language acquisition. *Studies in Second Language Acquisition, 13*, 57-72.

Gardner, R. C., Tremblay, P. F., & Masgoret, A. (1997). Towards a full model of second language learning: An empirical investigation. Modern Language Journal, 81(3) 344–362.

Gipps, Caroline (1994). Beyond Testing: Towards a Theory of Educational Assessment. London: The Falmer Press.

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning [Electronic Version]. *Learning and Teaching in Higher Education, 1*(1), 3-31.

Glasersfeld, E. (1995). *Radical constructivism: A way of knowing and learning*. London: The Falmer Press.

Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ*. New York: Bantam Books.

Goodrich, H. (1996). *Student self-assessment: At the intersection of metacognition and authentic assessment.* Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Goodrich, H. (1997). Understanding rubrics. Educational Leadership, 54(4), 14–17.

Graham, S. J. (2004). Giving up on modern foreign languages? Students' perceptions of learning French. Modern Language Journal, 88 (2) 171–191.

Gregory, K., Cameron, C., & Davies, A. (2000).*Self-assessment and goal-setting*. Merville, Canada: Connection.

Hamayan, E. (1995). Approaches to alternative assessment. *Annual Review of Applied Linguistics, 15*, 212-226.

Hanrahan, S. J. & Isaacs, G. (2001) Assessing self- and peer-assessment: the students' views, *Higher Educational Research and Development,* 20 (1), 53–70.

Harman, K., & Koohang, A. (2005). Discussion board: A learning object. *Interdisciplinary Journal of Knowledge & Learning Objects, 1*, 67-77. Retrieved from http://ijello.org/Volume1/v1p067-077Harman.pdf

Harris, M. & McCann, P. (1994). *Assessment.* London: Heinemann.

Hart, D. (1999) Opening assessment to our students, *Social Education*, 63(6), 343–345.

Heilenman, K. (1991). Self-assessment and placement: A review of the issues. In R. V. Teacher (Ed.), *Assessing foreign language proficiency of undergraduates.* Boston: Heinle & Heinle.

Henry, D. 1994. *Whole language students with low self-direction: A self-assessment tool.* Charlottesville: University of Virginia.

Higgs, Theodore V. & Clifford, Ray (1982). The push towards communication. In Theodore V. Higgs (Ed.), *Curriculum, competence, and the foreign language teacher* (pp. 57–79). Lincolnwood, IL: National Textbook Company.

Horner, S. & Shwery, C. (2002) Becoming an engaged, self-regulated reader, *Theory into Practice, 41*(2), 102–109.

Horwitz, E. K. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics, 21*, 112–126.

Horwitz, E.K., & Young, D.J. (Eds.). (1991). *Language Anxiety: From Theory and Research to Classroom Implications*. Englewood Cliffs, NJ: Prentice Hall.

Hughes, B., Sullivan, H. & Mosley, M. (1985). External evaluation, task difficulty, and continuing motivation. *Journal of Educational Research, 78*(4), 210-215.

Hung, D. (2001). Design principles for web-based learning; implications for Vygotskian thought. *Educational Technology, 41*(3), 33-41.

Hung, D., & Nichani, M. (2001). Constructivism and e-learning: balancing between the individual and social levels of cognition. *Educational Technology, 41*(2), 40-44.

Ingram, D. E. (1985). Assessing Proficiency: an over view of some aspect of testing. In K. Hyltenstam and m. Pienemann (Eds), Modeling and Assessing Second Language Acquisition (215-276). San Diego, CA: College-Hill Press.

Iran-Nejad, A. (1995). Constructivism as substitute for memorization in learning: Meaning is created by learner. *Education, 116*(1), 40-57.

Iwashita, Noriko, Brown, Annie, McNamara, Tim, & O'Hagan, Sally (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49.

Iwashita, Noriko. (2010). Features of Oral Proficiency in Task Performance by EFL and JFL Learners. In *Selected Proceedings of the 2008 Second Language Research Forum*, ed. Matthew T. Prior et al., 32-47. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2383.

Janssen-van Dieten, A. (1989). The development of a test of Dutch as a second language: The validity of self-assessment by inexperienced subjects. *Language Testing, 6*, 30-46.

Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *The School Psychology Review, 34*(1), 27-44.

Johansen, S. (2010). OPI Certification Process. Evaluation and Standardization Proficiency Standards Division, DLIFLC

Johnson, K. S. (2000). Affective component in the education of the gifted. *Gifted Child Today, 23*(4), 30-36.

Katz, E. (1994). *Self-concept and the gifted student.* Boulder, CO: Open Space.

Koohang, A., & Harman, K. (2005). Open source: A metaphor for e-learning. *Informing Science: The International Journal of an Emerging Trans discipline, 8*, 75-86.

Krashen, S. (1982). Principles and practice in second language acquisition. Oxford: Pergamon Institute of English.

Krashen, S.D., & Terrell, T.D. (1983). *The natural approach: Language acquisition in the classroom*. London: Prentice Hall Europe.

Krashen, S. (1985). *The input hypothesis: Issues and implications*. Torrance CA: Laredo Publishing Company, Inc.

Krausert, S. (1992). determining the usefulness of self-assessment of foreign language skills: Post-secondary ESL students' placement contribution. *Dissertation Abstract International, 52*, 3143A.

Kruse, E., & McKenna, S. (2008) U.S. House of Representatives. Armed Services Committee. Subcommittee on Oversight & Investigation. *Building language skills and culture competence in the military: DOD challenges in today's educational environment.*

Lantolf, J. P., & Thorne, S. L. (2006). *Socio-cultural theory and the genesis of second language development.* Oxford, United Kingdom: Oxford University Press.

Larsen –Freeman, Diane (2006). The emergence of complexity, fluency and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*(4), 590–619.

Leal, D. J. (2005-2006). The Word Writing CAFÉ: Assessing student writing for complexity, accuracy, and fluency. *The Reading Teacher, 59* (4), 340-349.

LeBlanc, R. & Painchaud, G (1985). Self-assessment as a second language placement instrument. TESOL Quarterly, 19, 73-87.

Lepkowski, W. (2006) Self-Assessment and Trained-Rater Assessment of Counselor Skills: Agreement Over Time and Gender Differences. *University of Nevada, Reno. ProQuest Dissertations and Theses*

Lewbel, S. R. & Hibbard, K. M. (2001). Are standards and true learning compatible? *Principal Leadership (High School Ed.),*1(5), 16–20.

Lim, D. H., & Morris, M. L. 2006. Influence of trainee characteristics, instructional satisfaction, and organizational climate on perceived learning and training transfer. *Human Resource Development Quarterly,* 17: 85–115.

Lockheed Martin (2010). Irregular warfare/stability operation capability portfolio. Focusing on the human dimension. Retrieved on 12/05/2012 , from http://www.fas.org/irp/program/collect/uas-army.pdf

Long, M. (1983). Native speaker/non-native speaker conversation and the negation of the comprehensive input. Applied Linguistics, 4, 126-141.

Lundsteen, W. (1979). LISTENING: ITS IMPACT ON READING AND THE OTHER LANGUAGE ARTS. Revised ed. Urbana, IL: National Council of Teachers of English and the ERIC Clearinghouse on Reading and Communication Skills, ED 169 537.

Luoma, S. & Tarnanen, M. (2003). Creating a self-rating instrument for second language writing: from idea to implementation. *Language Testing, 20 (4),* 440-465.

Mabe, P. & West, S. (1982). Validity of self-evaluation of ability: a review and Meta- Analysis: *Journal of Applied Psychology*, 67(3), 280-296.

Magnan, S. S. (1988). Grammar and the ACTFL oral proficiency interview: Discussion and data. *The Modern Language Journal, 72*(3), 266–276.

Marmolejo, F. (2010) Deficiency in Foreign Language Competency: What Is Wrong with the U.S. Educational System. *The Chronicle of Higher Education*

Marta, (2007). Immersion Education: *An Overview of Theory, Research and Practice.*

McDonald, B. (2007). Self Assessment for Understanding. *The journal of education*. The Trustees of Boston University, 181-1.

McFate, M. (2004). The Military Utility of Understanding Adversary Culture. Retrieved on January 10, 2011 from www.dtic.mil/doctrine/jel/jfq_pubs/1038.pdf.

McFate, M., & Jackson, A. (2005) An Organizational Solution for DOD's Cultural Knowledge Needs.

McMillan, J. H., & Hearn, J. (2009). Student self-assessment: The key to stronger student motivation and higher achievement. *The Education Digest, 74* (8), 39-44.

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52–75.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

Menard, S. (1995). Applied Logistic Regression Analysis. Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 106.

Mills, N., Pajares, F., & Herron, C. (2006). A reevaluation of the role of anxiety: Self-efficacy, anxiety, and their relation to reading and listening proficiency. Foreign Language Annals, 39(2) 276–295.

Mills, N., Pajares, F., & Herron, C. (2007). Self-efficacy of college intermediate French Students: Relation to achievement and motivation. Language Learning, 57(3) 417–442.

Moritz, C. E. B. (1995). *Self-assessment of foreign language proficiency: A critical analysis of issues and a study of cognitive orientations of French learners. ProQuest Dissertations and Theses.*

Moskal, B. M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research & Evaluation, 8*(14).

Moskal, B. M. (2010). "Self-Assessments: What are their Valid Uses?" Academy of Management Learning and Education, (9), 2.

Mowrer, C. (2006). Self –assessment and gender consideration in utilizing the CAFE (complexity, Accuracy, Fluency, and Evaluation) to assess students word writing ability. Ohio University. *ProQuest Dissertations and Theses.*

Murphy, J. (2005, February). Unpacking the foundations of the ISLLC Standards and addressing concerns in the academic community. *Educational Administration Quarterly, 41*(1), 154-191.

Nelson, J. R., Smith, D. J., & Colvin, G. 1995. The effects of a peer-mediated self-evaluation procedure on the recess behavior of students with behavioral problems. *Remedial and Special Education, 16*(2), 117-126.

Norris, John M. & Ortega, Lourdes (2003). Defining and measuring SLA. In Catherine Doughty & Michael H. Long (Eds.), *Handbook of second language acquisition* (pp. 716–761). London: Blackwell.

North, B. (1993). The development of descriptors scales of proficiency: Perspectives, problems, and possible methodology Washington DC: National Foreign Language Center.

Olsen, E. (2009). Irregular warfare/stability operation capability portfolio. Focusing on the Human dimension. Lockheed Martin (2010).

Omaggio, A. (1986). *Teaching language in context: Proficiency oriented instruction.* Boston, MA: Heinle and Heinle Publishers.

O'Malley, J. M., & Pierce, L. V. (1996). Authentic assessment for English language learners. Boston, MA: Addison-Wesley Publishing Company.

OPI 2000, (2010). *Tester Workshop Training Manual.* DLIFLC. Evaluation and Standardization Directorate. Proficiency Standards Division.

Ortega, Lourdes (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 4*(24), 492–518.

Oscarson, A. D. (2009) Self-Assessment of Writing in Learning English as a Foreign Language. *Goteborg Studies in Educational Science 277,* Goteborg University. (P.31-47).

Oscarson, M. (1977). *Self-assessment in Foreign Language Learning.* Strasbourg: Council of Europe.

Oscarson, M. (1978). *Approaches to self-assessment in foreign language learning.* Oxford, England: Pergamon Press.

Oscarson, M. (1980). Approaches to Self-Assessment in Foreign Language Learning. Oxford: Pergamon Press.

Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In: *Encyclopedia of language and education. –Vol.7: Language testing and assessment*. Dordrecht: Kluwer Academic, 175-187.

Pallier, G. (2003). Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles, 48*(5/6), 265-276.

Pajares, F., (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly, 19,* 139-158.

Piaget, J. (1953). *The Origins of Intelligence in Children*. London: Routledge and Kegan Paul.

Paris, S. G. & Paris, A. H. (2001). Classroom applications of research on self-regulated learning, *Educational Psychologist,* 36 (2), 89–101.

Parry, T. (2010). Private Conversation. Associate Provost of Evaluation and Standardization, DLIFLC. Monterey, CA.

Peterson, S. (1998). Evaluation and teachers' perception of gender in sixth grade student. *Research in the Teaching of English*, 33, 181-206.

Pierce, B. M., Swain, M., & Hart, D. (1993). Self-assessment, French immersion, and locus of control. *Applied Linguistics, 14,* 25-42.

Pintrich, P. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P.

Pintrich & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 452–502). San Diego, CA: Academic Press.

Popham, J. W. (1997). What's wrong-and what's right-with rubrics. *Educational Leadership, 55*(2), 72–75.

PRC, INC. (1997). *A guide for evaluating foreign language immersion training.* Reston, VA: PRC, INC

Quinlan, A. (2006). *Assessment made easy: Scoring rubrics for teachers from K–college.* Lanham, MD: Rowman and Littlefield Education.

Raymond, E. (2000). Cognitive Characteristics. *Learners with Mild Disabilities* (pp. 169-201). Needham Heights, MA: Allyn & Bacon, A Pearson Education Company.

Ridley, J. (2003). Learners' ability to reflect on language and on their learning. In D. Little, J. Ridley, & E. Ushioda (Eds.), *Learner autonomy in the foreign language classroom.* (pp. 78–89) Dublin: Authentik.

Rivers, W. (2001) Autonomy at all costs: an ethnography of metacognitive self-assessment and self-management among experienced language learners, *The Modern Language Journal,* 85(2), 279–290.

Ross, J. A., Rolheiser, C. & Hogaboam-Gray, A. (1998). Skills training versus action research in-service: Impact on student attitudes to self-evaluation. *Teaching and Teacher Education, 14*(5), 463-477.

Ross, J. A. Hogaboam-Gray, A. & Rolheiser, C. (2002). Self-evaluation in grade 11 mathematics: Effects on achievement and student beliefs about ability. In: D. McDougall. (Ed.), *OISE papers on mathematical education.* Toronto: University of Toronto.

Ross, J.A. & Starling, M. (2005). Achievement and self-efficacy effects of self-evaluation training in a computer-supported learning environment. *Paper presented at the annual meeting of the American Educational Research Association, Montreal.*

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. Language Testing, 15(1), 1-20.

Scarborough, R. (2009). Lack of Translators Hurts U.S. War on Terror. Student news daily. Washington Times.

Schunk, D. H. 1996. Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal, 33*(2), 359-382.

Schunk, D. (2003). Self-efficacy for reading and writing: Influence of modeling, goal-setting, and self-evaluation. *Reading & Writing Quarterly, 19*, 159–172.

Schwartz, M. (2009).*Congressional Research services (CRS). Department of Defense Contractors in Iraq and Afghanistan: Background and Analysis.*

Sekula, J., Buttery, T. & Guyton, E. (1996). Authentic assessment. In *Handbook of Research on Teacher Education.* New York: Prentice Hall International, 5-15.

Shadrick, S. B., & Schaefer, P. S. (2007). *Development and content validation of crisis response training package Red Cape: Crisis action planning and execution*. (Research Report 1875). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Shrauger, J. S. & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessment. *Psychological Bulletin, 90, 322-351.*

Sitzmann, T.; Ely, K.; Brown, K. Bauer, K. (2010) Self-Assessment of Knowledge: A Cognitive Learning or Affective Measure. *Academy of Management Learning & Education*, 2010, Vol. 9, No. 2, 169–191.

Sparks, G. E. 1991. *The effect of self-evaluation on musical achievement, attentiveness and attitudes of elementary school instrument students.* (Unpublished doctoral dissertation, Louisiana State University and Agricultural and Mechanical College).

Stallings, V. & Tascione, C. (1996). Student self-assessment and self-evaluation. *Mathematics Teacher, 89*, 548–55.

Stanton, H. E. (1976). Self-grading as an assessment method. *Improving College and University Teaching, 26*(4), 236-238.

Stein, M. (1999). Developing Oral Proficiency in the Immersion Classroom. *The Bridge: From Research to Practice.* ACIE Newsletter, Volume 2, Number 3.

Stiggins, R. J. (2001). *Student-involved classroom assessment* (3rd ed.) Upper Saddle River, NJ: Merrill/Prentice-Hall.

Strong-Krause, D. (1997, March). *How effective is self-assessment for ESL placement?* Paper presented at the annual meeting of TESOL, Orlando, FL.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235-256). Cambridge, MA: Newbury House Publishers.

Swanson, J. & Lease, S. (1990). Gender Differences in self-rating of abilities and skills. *Career Development Quarterly, 5* (1), 47-59.

Thurstone, L. L. (1982). Attitude can be measured. In B. D. Wright & G. Masters (EDS.), *Rating scale analysis: research measurement* (pp. 10-15). Chicago. Mesa Press.

Tompkins, G. E. (2004). *Language arts. Pattern of practice* (6th Ed.) Pearson.

Tremblay, P. F., & Gardner, R. C. (1995). Expanding the motivation construct in language learning. *Modern Language Journal, 79* (4), 505–520.

U.S. Commission Report on 9/11(2002). *Final report of the national commission on terrorist attacks upon the United States.*

Ushioda, E. (2003). Motivation as a socially mediated process. In D. Little, J. Ridley, & E. Ushioda (Eds.), *Learner autonomy in the foreign language classroom.* (pp. 93–129) Dublin: Authentik.

Vygotsky, L.S. (1962). *Thought and Language.* Cambridge, MA: MIT Press

Walczyk, J. J., & Hall, V. C. (1989). Effects of examples and embedded questions on the accuracy of comprehension self assessments. *Journal of Educational Psychology,* 81: 435–437.

Wesche, M. B., Morrison, F., Ready, D., & Pawley, C. (1990). French immersion: Postsecondary consequence for individuals and universities. *Canadian Modern Language Review, 46,* 430-51.

White, L. (1987). Against comprehensible input: The input hypothesis and the development of second language competence. *Applied Linguistics, 8*(2), 95-110.

Wiggins, G. (1998). *Educative assessment: designing assessments to inform and improve student performance* (San Francisco, CA, Jossey-Bass).

Wildes, K. (2005).U.S. Troops learn about Iraqi culture ahead of deployment. Retrieved on 1/8/2011 from http://www.voanews.com/english/news/a-13-2005-08-05-voa28.html.

Wilhelm, J. & Friedemann, P. (2002). *Hyper learning: Where Projects, inquiry, and technology meet. York, ME: Stenhouse.*

Williams, M., & Burden, R. (1997). *Psychology for language teachers. A social constructivist approach*. Cambridge: Cambridge University Press.

Wilson, K. M. (1999). Validity of global self-ratings of ESL speaking proficiency based on an FSI/ILR-?referenced scale. *Educational Testing Service.* RR-99-13.

Wolochuk, A. (2009). *Adult English Learners' self-assessment of second language proficiency: Contexts and conditions. ProQuest Dissertations and Theses.*

Wood, D., Bruner, J., & Ross, G. (1976). The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry, 17,* 89-100.

Yang, W., & Xu, X. Z. (2008). Self-assessment in second language learning. School of Foreign Languages, Shenzhen Polytechnic, Shenzhen 518055, China, US-China Foreign language, ISSN 1539-8080, USA, Volume 6, No. 5 (Serial no. 56).

Young, D. J. (1986). The relationship between anxiety and foreign language oral proficiency ratings. *Foreign Language Annals, 19,* 439– 445.

Young, D. J. (1990). An investigation of students' perspectives on anxiety and speaking. *Foreign Language Annals, 23,* 539–553.

Zimmerman, B. J. (1998). Academic studying and the development of personal skill: A self-regulatory perspective. *Educational Psychologist 33*(2-3), 73-86.

Zimmerman, B. J., & Schunk, D. H. (2001). Reflections on theories of self-regulated learning and academic achievement. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives (2nd ed.).* (pp. 289-307). Mahwah, N.J.: Lawrence Erlbaum Associates Publishers

Zimmerman, B., & Schunk, D. (2004). Self-regulating intellectual processes and outcomes: A social cognitive perspective. In D. Dai & R. Sternberg (Eds.), *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development 19* (pp. 323–349). Mahwah, NJ: Lawrence Erlbaum Associates.

APPENDICES

APPENDIX A

The Approval to Conduct the Study at DLIFLC

DEPARTMENT OF THE ARMY
DEFENSE LANGUAGE INSTITUTE FOREIGN LANGUAGE CENTER
AND PRESIDIO OF MONTEREY
MONTEREY CA 93944-3236

5 July 2011

Institutional Review Board
U.S. Army Assurance: DOD A20209

Institutional Review Board for the Protection of Human Subjects
Argosy University San Francisco Bay Area
1005 Atlantic Avenue
Alameda, CA 94501

Dear Dr. Griffith:

On behalf of the U.S. Army Defense Language Institute Foreign Language Center (DLIFLC), I am writing to formally indicate our awareness of a research project proposed by Mr. Salem Elfiky, a graduate student at Argosy University.

This research project, tentatively entitled *INVESTIGATING THE RELATIONSHIP BETWEEN STUDENTS' SELF-ASSESSMENT AND RATINGS OBTAINED FROM A FORMAL ORAL PROFICIENCY INTERVIEW (OPI),* has been reviewed by Dr. Donald Fischer (DLIFLC Provost), and he has approved the use of DoD personnel (military and civilian) as participants in this research project.

I have been informed that the Argosy IRB will conduct the review and maintain institutional oversight of this project. Once the Argosy IRB has completed its review of the project, I ask that a copy of the outcome of that review (and approval number) be send to me so we may maintain a folder on this project in our file of current research projects.

If you have any questions or concerns, please feel free to contact me.

Sincerely,

J. Jeffrey Crowson, Ph.D.
Senior Research Psychologist
IRB Chair
(831) 242-3788
jeff.crowson@us.army.mil

APPENDIX B

The Permission Letter to use Four Can-Do Items from DLIFLC Publication

Research and Analysis Division
Defense Language Institute Foreign Language Center
DoD Center Monterey Bay
400 Gigling Road
Seaside, CA 93955
8 July 2011

Mr. Salem Elfiky
Proficiency Standards Division
Defense Language Institute Foreign Language Center
DoD Center Monterey Bay
400 Gigling Road
Seaside, CA 93955

Dear Mr. Elfiky,

I am pleased to confirm in writing that you were granted permission by the Research and Analysis Division to incorporate in the self-assessment instrument that you developed for your doctoral dissertation, *Investigating the Relationship Between Students' Self-Assessment and Ratings Obtained From a Formal Oral Proficiency Interview (OPI)*, four self-assessment items included in the following DLIFLC publication: *A Guide for Evaluating Foreign Language Immersion Training* (Research and Analysis Division Research Report 97-01).

Yours sincerely,

Gordon L. Jackson

Gordon L. Jackson, Ph.D.
Senior Researcher
Research and Analysis Division

APPENDIX C

Interagency Language Roundtable Language Skill Level Descriptions (ILR)

Speaking

INTERAGENCY LANGUAGE ROUNDTABLE

LANGUAGE SKILL LEVEL DESCRIPTIONS

http://www.govtilr.org

SPEAKING

Preface

The following proficiency level descriptions characterize spoken language use. Each of the six "base levels" (coded 00, 10, 20, 30, 40, and 50) implies control of any previous "base level's" functions and accuracy. The "plus level" designation (coded 06, 16, 26, etc.) will be assigned when proficiency substantially exceeds one base skill level and does not fully meet the criteria for the next "base level." The "plus level" descriptions are therefore supplementary to the "base level" descriptions. A skill level is assigned to a person through an authorized language examination. Examiners assign a level on a variety of performance criteria exemplified in the descriptive statements. Therefore, the examples given here illustrate, but do not exhaustively describe, either the skills a person may possess or situations in which he/she may function effectively. Statements describing accuracy refer to typical stages in the development of competence in the most commonly taught languages in formal training programs. In other languages, emerging competence parallels these characterizations, but often with different details. Unless otherwise specified, the term "native speaker" refers to native speakers of a standard dialect. "Well-educated," in the context of these proficiency descriptions, does not necessarily imply formal higher education; however, in cultures where formal higher education is common, the language-use abilities of persons who have had such education is considered the standard. That is, such a person meets contemporary expectations for the formal, careful style of the language, as well as a range of less formal varieties of the language.

Speaking 0 (No Proficiency)

Unable to function in the spoken language. Oral production is limited to occasional isolated words. Has essentially no communicative ability. (Has been coded L-0 in some nonautomated applications. [Data Code 00]

Speaking 0+ (Memorized Proficiency)

Able to satisfy immediate needs using rehearsed utterances. Shows little real autonomy of expression, flexibility or spontaneity. Can ask questions or make statements with reasonable accuracy only with memorized utterances or formulae. Attempts at creating speech are usually unsuccessful.

Examples: The individual's vocabulary is usually limited to areas of immediate survival needs. Most utterances are telegraphic; that is, functors (linking words, markers and the like) are omitted, confused or distorted. An individual can usually differentiate most significant sounds when produced in isolation but, when combined in words or groups of words, errors may be frequent. Even with repetition, communication is severely limited even with people used to dealing with foreigners. Stress, intonation, tone, etc. are usually quite faulty. (Has been coded S-0+ in some nonautomated applications.) [Data Code 06]

Speaking 1 (Elementary Proficiency)

Able to satisfy minimum courtesy requirements and maintain very simple face-to-face conversations

on familiar topics. A native speaker must often use slowed speech, repetition, paraphrase, or a combination of these to be understood by this individual. Similarly, the native speaker must strain and employ real-world knowledge to understand even simple statements/questions from this individual. This speaker has a functional, but limited proficiency. Misunderstandings are frequent, but the individual is able to ask for help and to verify comprehension of native speech in face-to-face interaction. The individual is unable to produce continuous discourse except with rehearsed material.

Examples: Structural accuracy is likely to be random or severely limited. Time concepts are vague. Vocabulary is inaccurate, and its range is very narrow. The individual often speaks with great difficulty. By repeating, such speakers can make themselves understood to native speakers who are in regular contact with foreigners but there is little precision in the information conveyed. Needs, experience or training may vary greatly from individual to individual; for example, speakers at this level may have encountered quite different vocabulary areas. However, the individual can typically satisfy predictable, simple, personal and accommodation needs; can generally meet courtesy, introduction, and identification requirements; exchange greetings; elicit and provide, for example, predictable and skeletal biographical information. He/she might give information about business hours, explain routine procedures in a limited way. and state in a simple manner what actions will be taken. He/she is able to formulate some questions even in languages with complicated question constructions. Almost every utterance may be characterized by structural errors and errors in basic grammatical relations. Vocabulary is extremely limited and characteristically does not include modifiers. Pronunciation, stress, and intonation are generally poor, often heavily influenced by another language. Use of structure and vocabulary is highly imprecise. (Has been coded S-1 in some nonautomated applications.) [Data Code 10]

Speaking 1+ (Elementary Proficiency, Plus)

Can initiate and maintain predictable face-to-face conversations and satisfy limited social demands. He/she may, however, have little understanding of the social conventions of conversation. The interlocutor is generally required to strain and employ real-world knowledge to understand even some simple speech. The speaker at this level may hesitate and may have to change subjects due to lack of language resources. Range and control of the language are limited. Speech largely consists of a series of short, discrete utterances.

Examples: The individual is able to satisfy most travel and accommodation needs and a limited range of social demands beyond exchange of skeletal biographic information. Speaking ability may extend beyond immediate survival needs. Accuracy in basic grammatical relations is evident, although not consistent. May exhibit the more common forms of verb tenses, for example, but may make frequent errors in formation and selection. While some structures are established, errors occur in more complex patterns. The individual typically cannot sustain coherent structures in longer utterances or unfamiliar situations. Ability to describe and give precise information is limited. Person, space and time references are often used incorrectly. Pronunciation is understandable to natives used to dealing with foreigners. Can combine most significant sounds with reasonable comprehensibility, but has difficulty in producing certain sounds in certain positions or in certain combinations. Speech will usually be labored. Frequently has to repeat utterances to be understood by the general public. (Has been coded S-1+ in some nonautomated applications.) [Data Code 16]

Speaking 2 (Limited Working Proficiency)

Able to satisfy routine social demands and limited work requirements. Can handle routine work-

related interactions that are limited in scope. In more complex and sophisticated work-related tasks, language usage generally disturbs the native speaker. Can handle with confidence, but not with facility, most normal, high-frequency social conversational situations including extensive, but casual conversations about current events, as well as work, family, and autobiographical information. The individual can get the gist of most everyday conversations but has some difficulty understanding native speakers in situations that require specialized or sophisticated knowledge. The individual's utterances are minimally cohesive. Linguistic structure is usually not very elaborate and not thoroughly controlled; errors are frequent. Vocabulary use is appropriate for high-frequency utterances. but unusual or imprecise elsewhere.

Examples: While these interactions will vary widely from individual to individual, the individual can typically ask and answer predictable questions in the workplace and give straightforward instructions to subordinates. Additionally, the individual can participate in personal and accommodation-type interactions with elaboration and facility; that is, can give and understand complicated, detailed, and extensive directions and make non-routine changes in travel and accommodation arrangements. Simple structures and basic grammatical relations are typically controlled; however, there are areas of weakness. In the commonly taught languages, these may be simple markings such as plurals, articles, linking words, and negatives or more complex structures such as tense/aspect usage, case morphology, passive constructions, word order, and embedding. (Has been coded S-2 in some nonautomated applications.) [Data Code 20]

Speaking 2+ (Limited Working Proficiency, Plus)

Able to satisfy most work requirements with language usage that is often, but not always, acceptable and effective. The individual shows considerable ability to communicate effectively on topics relating to particular interests and special fields of competence. Often shows a high degree of fluency and ease of speech, yet when under tension or pressure, the ability to use the language effectively may deteriorate. Comprehension of normal native speech is typically nearly complete. The individual may miss cultural and local references and may require a native speaker to adjust to his/her limitations in some ways. Native speakers often perceive the individual's speech to contain awkward or inaccurate phrasing of ideas, mistaken time, space and person references, or to be in some way inappropriate, if not strictly incorrect.

Examples: Typically the individual can participate in most social, formal, and informal interactions, but limitations either in range of contexts, types of tasks or level of accuracy hinder effectiveness. The individual may be ill at ease with the use of the language either in social interaction or in speaking at length in professional contexts. He/she is generally strong in either structural precision or vocabulary, but not in both. Weakness or unevenness in one of the foregoing, or in pronunciation, occasionally results in miscommunication. Normally controls, but cannot always easily produce general vocabulary. Discourse is often incohesive. (Has been coded S-2+ in some nonautomated applications.) [Data Code 26]

Speaking 3 (General Professional Proficiency)

Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations in practical, social and professional topics. Nevertheless, the individual's limitations generally restrict the professional contexts of language use to matters of shared knowledge and/or international convention. Discourse is cohesive. The individual uses the language acceptably, but with

some noticeable imperfections; yet, errors virtually never interfere with understanding and rarely disturb the native speaker. The individual can effectively combine structure and vocabulary to convey his/her meaning accurately. The individual speaks readily and fills pauses suitably. In face-to-face conversation with natives speaking the standard dialect at a normal rate of speech, comprehension is quite complete. Although cultural references, proverbs and the implications of nuances and idiom may not be fully understood, the individual can easily repair the conversation. Pronunciation may be obviously foreign. Individual sounds are accurate: but stress, intonation and pitch control may be faulty.

Examples: Can typically discuss particular interests and special fields of competence with reasonable ease. Can use the language as part of normal professional duties such as answering objections, clarifying points, justifying decisions, understanding the essence of challenges, stating and defending policy, conducting meetings, delivering briefings, or other extended and elaborate informative monologues. Can reliably elicit information and informed opinion from native speakers. Structural inaccuracy is rarely the major cause of misunderstanding. Use of structural devices is flexible and elaborate. Without searching for words or phrases, the individual uses the language clearly and relatively naturally to elaborate concepts freely and make ideas easily understandable to native speakers. Errors occur in low-frequency and highly complex structures. (Has been coded S-3 in some nonautomated applications.) [Data Code 30]

Speaking 3+ (General Professional Proficiency, Plus)

Is often able to use the language to satisfy professional needs in a wide range of sophisticated and demanding tasks.

Examples: Despite obvious strengths, may exhibit some hesitancy, uncertainty, effort or errors which limit the range of language-use tasks that can be reliably performed. Typically there is particular strength in fluency and one or more, but not all, of the following: breadth of lexicon, including low- and medium-frequency items, especially socio-linguistic/cultural references and nuances of close synonyms; structural precision, with sophisticated features that are readily, accurately and appropriately controlled (such as complex modification and embedding in Indo-European languages); discourse competence in a wide range of contexts and tasks, often matching a native speaker's strategic and organizational abilities and expectations. Occasional patterned errors occur in low frequency and highly-complex structures. (Has been coded S-3+ in some nonautomated applications.) [Data Code 36]

Speaking 4 (Advanced Professional Proficiency)

Able to use the language fluently and accurately on all levels normally pertinent to professional needs. The individual's language usage and ability to function are fully successful. Organizes discourse well, using appropriate rhetorical speech devices, native cultural references and understanding. Language ability only rarely hinders him/her in performing any task requiring language; yet, the individual would seldom be perceived as a native. Speaks effortlessly and smoothly and is able to use the language with a high degree of effectiveness, reliability and precision for all representational purposes within the range of personal and professional experience and scope of responsibilities. Can serve as in informal interpreter in a range of unpredictable circumstances. Can perform extensive, sophisticated language tasks, encompassing most matters of interest to well-educated native speakers, including tasks which do not bear directly on a professional specialty.

Examples: Can discuss in detail concepts which are fundamentally different from those of the target culture and make those concepts clear and accessible to the native speaker. Similarly, the individual can understand the details and ramifications of concepts that are culturally or conceptually different from his/her own. Can set the tone of interpersonal official, semi-official and non-professional verbal exchanges with a representative range of native speakers (in a range of varied audiences, purposes, tasks and settings). Can play an effective role among native speakers in such contexts as conferences, lectures and debates on matters of disagreement. Can advocate a position at length, both formally and in chance encounters, using sophisticated verbal strategies. Understands and reliably produces shifts of both subject matter and tone. Can understand native speakers of the standard and other major dialects in essentially any face-to-face interaction. (Has been coded S-4 in some nonautomated applications.) [Data Code 40]

Speaking 4+ (Advanced Professional Proficiency, Plus)

Speaking proficiency is regularly superior in all respects, usually equivalent to that of a well educated, highly articulate native speaker. Language ability does not impede the performance of any language-use task. However, the individual would not necessarily be perceived as culturally native.

Examples: The individual organizes discourse well. employing functional rhetorical speech devices, native cultural references and understanding. Effectively applies a native speaker's social and circumstantial knowledge; however, cannot sustain that performance under all circumstances. While the individual has a wide range and control of structure, an occasional nonnative slip may occur. The individual has a sophisticated control of vocabulary and phrasing that is rarely imprecise, yet there are occasional weaknesses in idioms, colloquialisms, pronunciation, cultural reference or there may be an occasional failure to interact in a totally native manner. (Has been coded S-4+ in some nonautomated applications.) [Data Code 46]

Speaking 5 (Functionally Native Proficiency)

Speaking proficiency is functionally equivalent to that of a highly articulate well-educated native speaker and reflects the cultural standards of the country where the language is natively spoken. The individual uses the language with complete flexibility and intuition, so that speech on all levels is fully accepted by well-educated native speakers in all of its features, including breadth of vocabulary and idiom, colloquialisms and pertinent cultural references. Pronunciation is typically consistent with that of well-educated native speakers of a non-stigmatized dialect. (Has been coded S-5 in some nonautomated applications.) [Data Code 50]

APPENDIX D

Informed Consent Form for Survey Participants

INFORMED CONSENT FORM

TITLE OF STUDY

Relationship between self-assessment and formal assessment in speaking (OPI)

RESEARCHER

Mr. Salem Abdelhamid Elfiky

PURPOSE OF THE STUDY

The purpose of the study is to investigate the relationship between two types of measures of oral proficiency—level scores inferred from a self-assessment instrument and ratings obtained from a formal Oral Proficiency Interview (OPI). If you agree to participate in the study, 1) you will complete the self-assessment instrument, and your score on the instrument will be compared with your official, end-of-course OPI score; and 2) your DLPT listening and reading scores will be requested from DLI's Academic Records Office for use in future studies at DLI that examine the relationships among speaking, listening, and reading scores.

CONFIDENTIALITY

Confidentiality will be maintained in compliance with applicable federal regulations and statutes. The limits of confidentiality may be broken in the case of a court subpoena, other lawful means. Neither your name nor any other information that would identify you personally will be mentioned in any reports or presentations on the results of this study or of any future analyses involving your DLPT scores in speaking, listening and reading. If the researcher should mention the responses of an individual student, he will refer to the student by a random ID number assigned to maintain student anonymity.

WHAT YOU WILL DO

If you agree to participate in the study, you will complete a speaking self-assessment questionnaire which has two parts. The first part requests biographical information that is needed to examine the relationship between different variables and accuracy in self-assessment. In the second part you will indicate how easy/difficult it is for you to carry out given speaking

tasks in *real life situations* in the foreign language that you have been studying here at the DLIFLC. Participating in the study will take about 45 minutes. The self-assessment questionnaire will be administered about one week before your OPI, during a regular class period.

RISKS AND BENEFITS

Participation in the study is completely voluntary and there will be no negative consequences of any kind for not participating or from withdrawing from the study at any time. The researcher will be able to provide you with your survey score and you can use this information for your own language learning process. If you would like to know your score on the self-assessment instrument, please contact me.

MORE INFORMATION

If you have any questions or would like more information, please call me at 242-3761, send me an email message at salem.a.elfiky@us.army.mil, or stop by the DoD Center, 400 Gigling Road in Seaside, third floor. You may also contact my dissertation advisor at Argosy University, Dr. Scott Griffith, via email at slgriffith@argosy.edu.

I have read the above and I understand its contents and I agree to participate in the study.

PLEASE PRINT

_____

Last name First name Middle initial Date

_____

Signature

APPENDIX E

Can-Do Scale (CDS) Survey Instrument

Self-Assessment Survey of Speaking Proficiency

Can-Do-Scale (CDS)

The following self-assessment of speaking ability is intended to guide those who have not taken a U.S government-sponsored speaking test. It will produce an estimate of your speaking ability but is in no way a replacement for a formal, oral proficiency interview (OPI).

PLEASE PRINT

Examinee's name: Last_____ First_____ Class number: _____

1) Language tested: Arabic- Korean - Chinese (Circle one)

2) Education Level Completed: GED-HS- AA-BA-MA (Circle one)

3) Military branch: Army- Air Force- Navy- Marines (Circle one)

4) Rank: A) Officer B) Enlisted (Circle one)

5) Gender: Male – Female (Circle one)

6) Did you go on an in-country immersion? A) Yes B) No

   a. If yes, how long did you stay? _____

7) Age: A) (18- 20) B) (21-25) C) (26-30)

   D) (31-35) E) (36- 40) (Circle one)

8) Did you grow up speaking a language other than English? A) Yes B) No

   a. If yes, which language? _____

   b. At what age did you start speaking English? _____

9) Have you studied any other foreign language(s)?

   A) No B) Yes

a) If yes, which language(s) and how long?

#	Name of the language	Less than 6 months	6-12 months	1-2 years	More than 2 years
1					
2					
3					
4					

*Below is a series of statements describing speaking tasks that require use of spoken foreign language. Please read each statement carefully and check the appropriate box to indicate how well you can perform the task in the foreign language in **REAL-LIFE SITUATION** if you needed to do so. Please complete the survey by answering the 30 items.*

1. I can ask for directions on the street.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

2. I can order a simple meal.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

3. I can buy a needed item (such as a bus or train ticket, groceries, or clothing).

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

4. I can ask and answer questions about date and place of birth, nationality, marital status, occupation, etc.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

5. I can introduce myself in social situations, and use appropriate greetings and leave- taking expressions.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

6. I can arrange for a hotel room or taxi ride.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

7. I can answer simple questions about my present job, studies, or other major activities.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

8. I can talk about a regular hobby or a favorite activity in detail accurately and using appropriate vocabulary.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

9. I can give someone directions on how to get from here to other locations (such as, hotel, restaurant, or post office in detail).

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

10. I can talk about my future plans, (e.g. My plans for the next 5 years) using appropriate future tenses.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

11. I can give detailed instructions to another person (e. g. Steps on how to join the army).

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

12. I can tell a detailed story about something that happened in the recent past.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

13. I can report the facts of a recent news story or current event on a topic of interest to me.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

14. I can talk about my present or most recent job or activity in some detail using factual information but not technical language.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

15. I can give detailed information about such topics as my family, and my house or room.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

16. I can interview a job applicant, taking care of details such as salary, qualifications, hours, and specific duties or requirements.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

17. I can provide autobiographical details about myself, including immediate plans and preferences.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

18. I can make myself understood most of the time; when talking with native speakers not used to speaking with foreigners about work related requirements.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

19. I can take and give detailed messages over the telephone or leave a detailed message on voice mail.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

20. I can describe in detail a person or place that is very familiar to me.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

21. I can speak in detail to another person about my everyday activities.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

22. I can speak to a group of educated native speakers on a professional subject and be sure I am communicating what I want to, without reading from a prepared text.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

23. On a social occasion, I can defend personal opinions about social or cultural topics, such as the need for educational reform or obesity problems among children.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

24. I can debate and argue the economic and health care systems of my country.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

25. I can successfully and effectively use the language to cope with an unusual problem (such as an undeserved traffic ticket).

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

26. I can successfully use the language to resolve a serious social mistake made by a colleague or myself, such as an inappropriate use of a proverb/saying/cultural reference, an incorrect understanding of the meaning of and response to a proverb/saying or cultural reference, or addressing a new acquaintance in a formal setting in an inappropriately informal manner.

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

27. I can successfully use the language to speculate about societal topics (such as, what the situation in the United States/global community might be if the former Soviet Union had not collapsed/ if 911 had not happened/if environmental protection laws were not in place).

☐ Quite Easily ☐ Easily ☐ With some difficulty ☐ With great difficulty ☐ Not at all

Please check the appropriate box to indicate the extent to which you agree or disagree with the following statements.

28. There are only a few grammatical features of the language that I try to avoid when discussing political or social issues.

☐ Strongly Agree ☐ Agree ☐ Neither agree nor disagree ☐ Disagree ☐ Strongly Disagree

29. I rarely feel unable to finish expressing a thought or idea when discussing social or political topics because of language limitations (grammar or abstract vocabulary).

☐ Strongly Agree ☐ Agree ☐ Neither agree nor disagree ☐ Disagree ☐ Strongly Disagree

30. I find it easy to follow and participate in conversations on societal or political topics among native speakers.

☐ Strongly Agree ☐ Agree ☐ Neither agree nor disagree ☐ Disagree ☐ Strongly Disagree

Thank you

APPENDIX F

DLIFLC Inter-Rater Reliability FY 2011

Hi Salem,

I hope you had a wonderful weekend.

Per your request, the following table illustrates the inter-rater reliability for Arabic MSA, Korean, and Chinese Mandarin languages FY2011:

FY11	Agreed with Final Score	Total Initial Ratings	Percentage
Arabic MSA	1261	1315	95.89
Chinese Mandarin	721	733	98.36
Korean	479	496	96.57

Should you have further questions, please don't hesitate to ask

V/R
Osama Abushariefeh
Programmer Analyst
DoDCM-Automated Systems
400 Gigling Road
Seaside, CA 93955
T: 831.242. 5425

Classification: UNCLASSIFIED
Caveats: FOUO